Statistical Methods For Biomarker Threshold Models in Clinical Trials

by

Parisa Gavanji

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
January 2017

Copyright © Parisa Gavanji, 2017
Abstract

In clinical trials, the main objective is to investigate the treatment effects on patients. However, many molecularly targeted drugs or treatments tend to benefit a subset of patients more, identified by a certain biomarker. The cut-point value defining patient subsets is often unknown. For this situation, the ordinary likelihood ratio test cannot be applied for testing treatment-biomarker interaction because of the model irregularities.

We develop a residual bootstrap method to approximate the distribution of a proposed test statistic to test for treatment-biomarker interaction in survival data. Simulation studies show that the residual bootstrap test works well. The proposed method is applied to BIG 1-98 randomized clinical trial of breast cancer with Ki-67 as biomarker to consider the treatment effects on patients in two subsets. We also extend the residual bootstrap method to clustered survival data with an application to data from the I-SPY 1 clinical trial with the estrogen receptor total score as a biomarker.

Another research topic of the thesis is deriving the asymptotic distribution of a penalized likelihood ratio test statistic for testing biomarker effect and treatment-biomarker interaction in binary data. The model can be viewed as a mixture of logistic regression models with unknown cut-point for which the regularity conditions
of ordinary likelihood methods are not satisfied. We first approximate the indicator
function defining biomarker subgroups by a smooth continuous function. To overcome
irregularities, we develop a penalized likelihood method, introducing a new idea of
using random penalty term. Proposing a new set of regularity conditions helps us to
study the properties and limiting distributions of the maximum penalized likelihood
estimates of the parameters. We further prove that the penalized likelihood ratio
test statistic has an asymptotic $\chi^2_3$ distribution under the null hypothesis. Extensive
simulation studies show that the proposed test procedure works well for hypothesis
testing. The proposed method is applied to a clinical trial of prostate cancer with the
serum pro-static acid phosphatase (AP) as a biomarker.
Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisors Prof. Wenyu Jiang and Prof. Bingshu Chen for their support, encouragement and guidance during my four-year Ph.D. study. I have also been lucky enough to benefit from their experiences in computation and real-data applications, stimulating conversations and directions for the new research problems. I would like to express my sincere appreciation to my examination committee members: Prof. Glen Takahara and Prof. Dongsheng Tu. Without the helpful advice from them, this thesis would not have been nearly as good. Many thanks to Prof. Xin Gao from York university for her time and efforts in serving as the external examiner. I am so indebted to my family and my parents Firouzan Gavanji and Fereshteh Shahsavari, for their patience, encouragement, understanding and love throughout my whole life. Last but not least, I am so deeply grateful for my husband Amin Tavakoli. Thank you for being there for me from the very beginning.
Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work and that any ideas or quotations from the work of other people, published or unpublished, are fully acknowledged in accordance with the standard referencing practices of the discipline.
Contents

Abstract i

Acknowledgments iii

Statement of Originality iv

Contents v

List of Tables viii

List of Figures ix

Chapter 1: Introduction 1
  1.1 Overview ................................................................. 1
  1.2 Statistical models .................................................. 4
    1.2.1 Cox proportional hazards models for survival data ........ 4
    1.2.2 Frailty model for clustered survival outcome data .......... 6
    1.2.3 Residual bootstrap method for linear regression model ...... 10
    1.2.4 Logistic regression model for binary outcome data .......... 13
    1.2.5 Penalized likelihood ratio test for mixture models .......... 14
  1.3 Tests for treatment-biomarker interaction .......................... 18
  1.4 Applications to clinical trials .................................... 19
    1.4.1 Breast cancer data with Ki-67 as biomarker ................ 19
    1.4.2 I-SPY 1 data with estrogen receptor total score (ER-TS) as
        biomarker ........................................................... 20
    1.4.3 Prostate cancer data with acid phosphatase (AP) as biomarker 22
  1.5 Organization of thesis ................................................. 23

Chapter 2: Residual bootstrap method for biomarker threshold models 24
  2.1 Overview ................................................................. 24
  2.2 Model and methods .................................................. 27
2.2.1 Model .................................................. 27
2.2.2 Residual bootstrap method ................................ 29
2.2.3 Permutation test ........................................ 32
2.3 Simulation studies ...................................... 33
  2.3.1 Simulation model design ................................ 33
  2.3.2 Simulation results .................................... 34
2.4 Application to breast international group (BIG) 1-98 data set ..... 40
2.5 Discussions .............................................. 46

Chapter 3: Residual bootstrap method for clustered survival outcome 48
  3.1 Overview .................................................. 48
  3.2 Biomarker threshold model with random effects ................. 51
  3.3 The residual bootstrap method ................................ 56
  3.4 Simulation studies ........................................ 59
    3.4.1 Simulation model design ................................ 59
    3.4.2 Simulation results .................................... 61
  3.5 Application to I-SPY 1 data set ............................ 67
  3.6 Discussions .............................................. 70

Chapter 4: Penalized likelihood ratio test for binary outcome data 72
  4.1 Overview .................................................. 72
  4.2 Model and notation ...................................... 74
  4.3 Penalized likelihood ratio test ............................ 77
    4.3.1 Regularity conditions .................................. 80
    4.3.2 Properties of the maximum penalized likelihood estimates . 82
    4.3.3 Asymptotic distribution of the maximum penalized likelihood estimate \( \hat{\theta} \) ......................... 87
    4.3.4 Asymptotic null distribution of the penalized likelihood ratio test statistic \( R_n \) ......................... 89
    4.3.5 Choice of penalty term ................................ 94
  4.4 Simulation studies ........................................ 94
    4.4.1 Simulation model design ................................ 94
    4.4.2 Simulations results .................................... 98
  4.5 Application to prostate cancer data set ........................ 105
  4.6 Discussions .............................................. 107

Chapter 5: Summary and conclusions 109
  5.1 Summary .................................................. 109
  5.2 Future work ............................................. 110
# List of Tables

2.1 The empirical size of the test, $H_0 : \beta_3 = 0$. ........................................ 36

2.2 The empirical power of RBT, $H_1 : \beta_3 \neq 0$. ................................. 38

2.3 Empirical bias and standard error of the profile likelihood estimate of $c$. 39

3.1 The empirical size of the test for testing $H_0 : \beta_3 = 0$ for clustered survival outcome. ................................................................. 62

3.2 The empirical power of RBT under $H_1 : \beta_3 \neq 0$, for clustered survival outcome with $m = 50$. .................................................. 64

3.3 The empirical power of RBT under $H_1 : \beta_3 \neq 0$, for clustered survival outcome with $m = 100$. .................................................. 65

3.4 Empirical bias and standard error of the profile likelihood estimate of $c$ for clustered survival outcome. ............................... 66

4.1 Empirical test sizes for testing $H_0 : \beta_2 = \beta_3 = 0$. ......................... 99

4.2 Empirical test sizes for testing $H_0 : \beta_2 = \beta_3 = 0$. ......................... 100

4.3 Empirical power of the test under $H_1 : \text{at least one of } \beta_2 \text{ or } \beta_3 \text{ is not equal to 0, with } c = 0.5$. ............................... 102

4.4 Empirical power of the test under $H_1 : \text{at least one of } \beta_2 \text{ or } \beta_3 \text{ is not equal to 0, with } c = 0.75$. ............................... 103
List of Figures

2.1 Histogram of the test statistic values ........................................ 42
2.2 Profile partial likelihood $\ell(\hat{\beta}_c, c)$ versus $c$ ........................ 44
2.3 K-M curves for DFS, for BIG 1-98 data set ................................. 45

3.1 Histogram of the three available biomarkers for I-SPY 1 data set ... 68
3.2 Profile integrated partial likelihood function vs. $c$ values, under $H_1$, for I-SPY 1 data set .......................................................... 69
3.3 K-M curves for DFS, for I-SPY 1 data set .................................... 70

4.1 Compare $w(x, c)$ to $I(x > c)$ for different $K$ values .................. 76
4.2 Q-Q plot for $R_n$ with asymptotic $\chi^2_3$ distribution .................. 101
4.3 Penalized likelihood ratio test statistic versus cut-point value, using $I(x > c), w(x, c)$ ................................................................. 104
Chapter 1

Introduction

1.1 Overview

In clinical trials, traditional randomized designs and broad-eligibility designs are often used to evaluate the treatment effects. However, there are many new treatments that potentially benefit a subset of the population more. For example, in a recent study conducted by the Canadian Cancer Trials Group (CCTG), Cetuximab was shown to improve the overall survival of Colorectal Cancer patients ($HR = 0.77$, 95% confidence interval: 0.64-0.92, $p = 0.005$, where $HR$ is the hazard ratio of the new treatment versus control treatment, Jonker, et al (2007) [40]). Further subset analysis showed that the Cetuximab treatment effect was mainly on a subset of the patients with wild-type K-ras tumour ($HR = 0.55, p < 0.001$), but not on the subset of patients with K-ras mutation tumours ($HR = 0.98, p = 0.89$) (Karapetis, et al (2008) [43]). In these situations, these conventional clinical designs are not sufficient to identify the new treatment effect, and they may lead to missing effective treatments because of inclusion of many patients that may not benefit from the new treatment or suffer from unnecessary toxicities.
Therefore, we are interested in identifying the subset of the patients that benefit more from the new treatment. Moreover, identifying the subset of the patients that may not benefit or benefit less from the new treatment can help to avoid unnecessary therapy. In the conventional designs, we focus on the effect of the new treatment on all patients, which may result in failure to detect the true treatment effect when the treatment benefits a subset of patients more.

In some biomedical studies, there are reliable biomarkers based on which we can identify different subsets of the patients. A biomarker is defined as a baseline patient characteristic that affects a patient’s response to certain treatment. In general, a biomarker is measured on a continuous scale. There are two different types of biomarkers: prognostic and predictive biomarkers (Hilsenbeck (2014) [37] and Clark (2008) [18]). A prognostic biomarker relates to the natural history of a disease, indicating the likely course of the disease in an untreated individual. For example, a prognostic biomarker identifies patients who will relapse and experience recurrence of their cancer disease regardless of the treatment they received. A predictive biomarker is defined as an indicator to identify sub-populations of patients who are most likely to respond to a certain treatment. Since a prognostic biomarker does not have interaction with treatment, while predictive biomarkers have significant interaction with the treatment, it is critical to distinguish these two types of biomarkers. In this thesis, we will focus on statistical methodologies for the predictive biomarkers.

In the presence of a biomarker, targeted randomized designs can be used to investigate the effect of the new treatment on the patients that may benefit more from the new treatment. Targeted randomized designs identify this subset at first, and then assign these patients to the new treatment and control arms randomly. Therefore,
targeted randomized designs reduce the number of patients required for the study in the presence of a biomarker. However, a reliable biomarker to identify these two subsets may not always exist. Freidlin, et. al. (2005) proposed an adaptive signature design for randomized clinical trials in the case that a gene-expression-based classifier is used to identify patients more sensitive to the new treatment, along with a test for overall treatment effect [32].

In other situations, a biomarker is available, but the cut-point value to identify these two subsets of the patients has not been identified. Jiang, et. al. (2007) proposed a biomarker-adaptive threshold design [39]. Not only does the proposed method provide a test for the treatment effect, but also it provides a validation of a cut-point for a pre-specified biomarker in order to identify a sensitive subset containing the patients that benefit from the new treatment. Also, the method can be used to test if the new treatment benefits all patients or a subset of patients defined by the biomarker. Chen, et. al. (2014) developed a hierarchical Bayesian method to make statistical inference on both cut-point and the treatment effect on the more sensitive subset defined by the pre-specified biomarker [11]. They treated the cut-point value as a random variable with a certain prior distribution.

One of the most important aspects of the biomarker study is testing the interaction between the treatment and the biomarker because it describes the different responses to the new treatment among different subsets defined by the biomarker (Wang, et. al. (2007), Werft, et. al. (2012) [69], [70]).

There has been a lot of research in the literature about testing the biomarker-treatment interaction. In most of them, the Cox’s proportional hazards model is considered (Cox, (1972) [21]). For example, Gray(1992) used splines in additive
models for the analysis of survival data to make statistical inference on covariate effects and their interactions [35]. Bonetti and Gelber (2000) used a subpopulation treatment effect pattern plot to consider the treatment-biomarker interaction in the proportional hazards model [9]. Also, fractional polynomials were used by Royston and Sauerbrei (2004) to model the interaction effect between the biomarker and the treatment [57]. Royston et. al. (2006) suggested that the biomarker should be treated as a continuous variable [56]. They did that for using all potential information given by the data.

However, in practical applications, a cut-point is useful to guide treatment decision. Therefore, it is important to identify the cut-point value. The usual approach for identifying the cut-point value is to consider a grid of possible points and choose the value that maximizes the likelihood function (Jiang, et. al. (2007) [39]). It is proved by Farragi and Simon (1996) that this approach may overestimate the true treatment effect in the subsets [28]. They proposed using a two-fold cross validation method to estimate and test the subset treatment effect. Also, Wacholder et al. (2010) developed a permutation test for the treatment-biomarker interaction in the case that sensitive subsets are defined by two thresholds [68].

1.2 Statistical models

1.2.1 Cox proportional hazards models for survival data

Measurement of time to the occurrence of a certain event is an important endpoint in cancer clinical trials. Suppose we are interested in testing the interaction between the new treatment and the biomarker with an unknown cut-point for time to event data. One of the most important models to be considered is Cox’s proportional hazards
model (Cox (1972) [21]). This model is used to investigate the association between survival time data as response variable and the treatment and the biomarker or other covariates as predictors. We assign treatments to the patients randomly, and consider time to event outcomes.

Survival analysis is an important topic in statistics, and contains methods for analyzing time to event data. In clinical trials, the event can be death, occurrence of a disease, etc. Some studies may consider instead a binary variable that shows death/alive as a response variable. In survival analysis, the response variable is defined as time to a certain event, which is a continuous variable. A characteristic of survival data is that some data are censored. Censoring occurs when the time to event is only partially available. In the usual “right censoring” cases, all that is known is that the time exceeds some observed values. Some reasons of right-censoring are loss to follow-up, drop out or study termination. For example, a patient who does not experience the event of interest during the study period is said to be right-censored (Lawless (2002)[44]).

Let \( T_i, C_i \) be the potential failure time and censoring time for patient \( i \), respectively. Let \( t_i = \min(T_i, C_i) \) be the observed failure or censoring time, and \( \delta_i = I(T_i < C_i) \) indicate the failure/censoring status. Let

\[
S(t) = P(T > t)
\]

be the survival function for \( T \). To model time-to-event data of this type, we define a hazard function as,

\[
\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)\Delta t}.
\]  (1.1)
The hazard function is roughly the likelihood that an event occurs in interval \([t, t+\Delta t)\), given \(\{T \geq t\}\).

To study the relationship between explanatory variables and failure time, Cox (1972, 1975) proposed a proportional hazards regression model with hazard function \([21, 20]\),

\[
\lambda(t) = \lambda_0(t)g(x, \beta),
\]

(1.2)

where \(x\) is a vector of explanatory variables, \(\beta\) is a vector of regression parameters, and \(\lambda_0(t)\) is a baseline hazard function. Here, we only consider the typical relative risk function \(g(x, \beta) = \exp(x'\beta)\). Let \(t_1, \ldots, t_n\) be the observed failure or censoring time for \(n\) patients. The parameter \(\beta\) can be estimated by maximizing the partial likelihood function defined as (Cox (1972), [21]),

\[
L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(x_i'\beta)}{\sum_{j=1}^{n} Y_j(t_i) \exp(x_j'\beta)} \right\}^{\delta_i},
\]

(1.3)

where \(Y_j(t)\) describes the risk set, \(Y_j(t) = 1\) if subject \(j\) is still under observation at time \(t\) and 0 otherwise.

In this thesis, we will study the treatment-biomarker interaction for time-to-event data in chapter 2. The statistical inference will be based on the partial likelihood for the Cox proportional hazards model.

### 1.2.2 Frailty model for clustered survival outcome data

In cancer clinical trials, large sample size is required to ensure adequate statistical power to detect some small but clinically meaningful treatment effects. The need for
a large sample of observations to have enough participants results in having observations located in different clusters. In this situation, clusters are independent, but the observations within a cluster may be dependent. Frailty models are often used to describe the dependence of observations within a cluster and the heterogeneity between clusters.

In survival analysis with clustered data, there are two different approaches to model the correlation within a cluster: the conditional methods and the marginal methods. A mixed effects model, belonging to the conditional models family, is defined as a model in which the mean response depends on both covariates and a vector of random effects. Frailty models are a special type of mixed effect model and belong to the conditional models family. Frailty models for proportional hazards models can have a multiplicative frailty factor. The hazard factor in the model can depend on covariates and can be modeled in either a parametric or a semi parametric way.

On the other hand, marginal models are models in which the mean response depends only on the covariates of interest, not depending on the vector of random effects or repeated measures. Actually, marginal models do not require distributional assumptions for the response variables, and provides a unified method to analyze different types of longitudinal data and clustered data. To estimate unknown parameters without distributional assumptions for the response variable, one can use Generalized Estimating Equations (GEE). In GEE, the correlation between repeated measurements or the observations within a cluster will be considered as a part of variance. Liang and Zeger (1986) [46] introduced a standard iterative algorithm to solve the GEE and estimate parameters. Also, they showed that when the marginal
mean has been correctly specified by covariates and regularity conditions hold, the estimator is consistent and asymptotically normally distributed.

The frailty factor is random and therefore a frailty distribution needs to be specified in the frailty model (Duchateau and Janssen (2008) [24]). Different choices of frailty distributions are studied in the literature by several authors. Vaupel, et. al. (1979), Clayton and Cuzick (1985), and Andersen, et al. (1993) [66, 19, 3] discussed the frailty model with a frailty term following a Gamma distribution. However, Gamma frailty models can only model positive association among survival outcomes. That is the reason that other frailty distributions like multivariate normal, inverse Gaussian, and log-normal were considered by McGilchrist and Aisbett (1991), Hougaard (1995), and Ripatti and Palmgren (2000) [52, 38, 55]. Using a normal distribution for the frailty term, we can handle negative dependencies.

Suppose that the observations are located in $m$ clusters, and we have $n_i$ observations in each cluster, for $i = 1, \ldots, m$. Also, $\delta_{ij} = I(\tilde{T}_{ij} < C_{ij})$ represents the survival status indicator, in which $\tilde{T}_{ij}$ is the true failure time, and $C_{ij}$ is the censoring time for the $j^{th}$ patient in the $i^{th}$ cluster. Let $t_{ij} = \min(\tilde{T}_{ij}, C_{ij})$ be the observed failure/censoring time, for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. A Cox frailty model with normal distribution for the frailty term can be written as

$$
\lambda_{ij}(t | b_i) = \lambda_0(t) \exp(X_{ij} \beta + b_i),
$$

(1.4)

where $\lambda_0(t)$ is the baseline hazard function, and $X_{ij}$ is a $1 \times p$ design vector for fixed effects. Also, $\beta$ is a $p \times 1$ vector of fixed-effect coefficients and $b_i$ is a random effect coefficient for the $i^{th}$ cluster. The random effect vector $b = (b_1, \ldots, b_m)^T$ is assumed to follow a multivariate normal distribution with mean $0$ and a variance matrix $\Sigma$ which
depends on a vector of parameters $\theta$ (Therneau and Grambsch (2000) [64]). Assume that $b_1, \ldots, b_m$ are i.i.d; therefore, $\Sigma(\theta) = \sigma^2 I_{m \times m}$, where $I$ is identity matrix.

We can integrate out the random effects to create the integrated partial likelihood as

$$IPL(\beta, \theta) = \frac{1}{(2\pi)^{m/2} \sigma^2} \int \prod_{i=1}^{m} PL_i(\beta, b) \exp(-\frac{b^T b}{2\sigma^2}) db,$$

where $m$ is the number of random effects, and $PL_i(\beta, b)$ is the usual Cox partial likelihood function for any fixed values of $\beta$ in the $i^{th}$ cluster with random effect $b_i$,

$$PL_i(\beta, b) = \prod_{j=1}^{n_i} \left\{ \frac{\exp(X_{ij} \beta + b_i)}{\sum_{k=1}^{m} \sum_{h=1}^{n_k} Y_{kh}(t_{ij}) \exp(X_{kh} \beta + b_k)} \right\}^{\delta_{ij}},$$

where $Y_{ij}(t)$ describes the risk set, $Y_{ij}(t) = 1$ if subject $j$ in the cluster $i$ is still under observation at time $t$ and 0 otherwise. Based on the results in Ripatti and Palmgren (2000), the integrated partial likelihood can be viewed as a likelihood function [55].

To obtain the maximum likelihood estimates of the parameters $\beta, b$, and $\theta$, we have to maximize the logarithm of the integrated partial likelihood $IPL$ with respect to the parameters. However, the integral in (1.5) cannot be solved easily. Ripatti and Palmgren (2000) applied the Laplace approximation to the integral in the integrated partial likelihood $IPL$ [55].

In this thesis, we will develop a test procedure for treatment-biomarker interaction for clustered survival data. The proposed method will be based on the maximum integrated partial likelihood with Laplace approximation. We will use the R-package “coxme” (developed by Therneau (2015) [63]) to estimate parameter $\beta$ and the cluster-specified random effect $b_i$. 
1.2.3 Residual bootstrap method for linear regression model

Our goal in this research is biomarker study. As we mentioned before, testing the
treatment-biomarker interaction is an interesting aspect of the biomarker study. Some
resampling methods are very useful to generate the distribution of the test statistic
in order to calculate the p-value for the test.

The particular goal of the bootstrap theory is a computer-based implementation
of basic statistical concepts. A great advantage of bootstrap is its simplicity. It is a
method to obtain standard errors and confidence intervals in complex estimation of
complex parameters. Bootstrap is also an appropriate way to control and check the
stability of the results. When the distribution of the statistic of interest is compli-
cated or unknown, or the sample size is not sufficient for making statistical inference,
bootstrapping is useful (Efron and Tibshirani (1986, 1993), [25, 26]).

Suppose we observe independent data points, \( x_1, \ldots, x_n \), denoted by \( x = (x_1, \ldots, x_n) \),
coming from a population distribution with unknown parameter, \( \theta \). To make statisti-
cal inference about \( \theta \), we compute a statistic of interest, \( T(x) \). For example, suppose
we are interested in doing hypothesis testing, and let \( \hat{T} \) be the observed value of a
test statistic \( T \). The p-value for the test can be calculated by,

\[
p = 1 - F_0(\hat{T}),
\]

where \( F_0(T) \) is the cumulative distribution function of \( T \) under the null hypothesis.
If we know \( F_0(T) \), then we can simply find the p-value. However, in many cases, the
distribution of the test statistic is not easy to obtain, for example, in the study of the
biomarker cut-point problem. We can apply bootstrap resampling methods in these
situations.

For the purpose of hypothesis testing in a parametric way, we must generate $B$ bootstrap samples from the null distribution. Having $\hat{T}^{*b}$ for each bootstrap sample, the p-value can be estimated as,

$$\hat{p} = \frac{\#\{\hat{T}^{*b} > \hat{T}\}}{B},$$

where each bootstrap sample is generated from the null distribution. To generate bootstrap samples from the null distribution, we can use a parametric bootstrap method such as residual bootstrap. In this part, we introduce the residual bootstrap method for linear regression models. The method is based on fitting the model to generate the original sample of residuals, and then the bootstrap sample is generated under the null distribution.

A linear regression model has the form $y_i = x_i^T \beta + \epsilon_i, \ i = 1, \ldots, n$, where $x_i$ is a $p \times 1$ vector of predictors, and $\beta$ is a $p \times 1$ vector of unknown regression coefficients. The error terms, $\epsilon_i$, are assumed to be a random sample from an unknown distribution $F$, having expectation 0. We are interested in estimating the regression coefficient vector $\beta$ from the observed data $(x_i, y_i), i = 1, \ldots, n$, and making statistical inference about it.

Efron and Tibshirani (1993) introduced the residual bootstrap method for linear regression models [26]. They showed that the bootstrap variance estimates for regression parameters are the same as $\text{var}(\hat{\beta}_j) = \sigma_F^2 (X^T X)^{-1}_{jj}, j = 1, \ldots, p$, where $\sigma_F^2$ is the variance of $\epsilon_i$, estimated by mean square error. Therefore, the bootstrap method gives us reasonable answers. Also, we can apply bootstrap methods to more general regression models, where the regression function is non-linear in parameter $\beta$. 
1.2. STATISTICAL MODELS

The idea behind the residual bootstrap method is that the explanatory variables in the regression models are fixed, and do not need to be resampled, and the error terms are i.i.d. The residual bootstrap for testing the null hypothesis $H_0: \beta_j = \beta_{j0}$ for some $j \in \{0, 1, \ldots, p\}$ in a linear regression model can be done using the test statistic $\hat{T} = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\text{var}(\hat{\beta}_j)}}$ in the following steps:

1) Fit the regression model to estimate $\beta$ and calculate residuals,

$$\hat{\epsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_j x_{ij} + \ldots + \hat{\beta}_p x_{ip}),$$

for $i = 1, \ldots, n$. An obvious estimate of $F$ is the empirical distribution of the $\hat{\epsilon}_i$. The empirical distribution, $\hat{F}$, assigns probability $\frac{1}{n}$ to each $\hat{\epsilon}_i$, $i = 1, \ldots, n$.

2) To generate bootstrap data, take a random sample of error terms, $\epsilon^*_1, \ldots, \epsilon^*_n$, by sampling with replacement from $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$. Then, the bootstrap response, $y^*_i$, is given by

$$y^*_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \ldots + \tilde{\beta}_{j-1} x_{i,j-1} + \beta_{j0} x_{ij} + \tilde{\beta}_{j+1} x_{i,j+1} + \ldots + \tilde{\beta}_p x_{ip} + \epsilon^*_i,$$

which is generated under the null hypothesis, and $\tilde{\beta}_j$'s, $j = 1, \ldots, p$, are estimated under the null hypothesis.

3) Use the bootstrapped data, $(x_i, y_i^*)$, to fit a regression model, and calculate the bootstrap version of the test statistic, $\hat{T}^*$.  

4) Repeat steps 2, 3, $B$ times to get $\hat{T}^*_1, \ldots, \hat{T}^*_B$. Use these values to calculate p-value for testing the null hypothesis by,

$$\hat{p} = \frac{\#\{\hat{T}^*_b > \hat{T}\}}{B}.$$
1.2. STATISTICAL MODELS

In this thesis, a similar idea will be applied to test statistics for time-to-event data, in Cox proportional hazards regression model and frailty model. Martingale or Cox-Snell residuals will be used to construct bootstrap samples for survival time and censoring indicator pairs for the statistical inference in chapter 2 and chapter 3.

1.2.4 Logistic regression model for binary outcome data

In some application scenarios, the response variable is a binary outcome which represents death/alive, and we are interested in testing the treatment-biomarker interaction effect for this type of data. The logistic regression model can be used to associate the response variable and the treatment and the biomarker as covariates. In this section, we will review some concepts of the logistic regression model. The logistic regression model is a special case of the generalized linear models.

“Generalized linear models (GLM) extend ordinary regression models to encompass non-normal response distributions by modeling functions of the mean response. A generalized linear model has three components. A random component which is the response variable $Y$ and its probability distribution, a systematic component which is explanatory variables used in a linear predictor function, and a link function that relates random component and systematic component” (Agresti (2013) [2]).

Let $Y_1, \ldots, Y_n$ be the response variables from the exponential dispersion family whose probability mass/density function can be written in the following form,

$$f(y_i, \theta_i) = a(\theta_i)b(y_i)\exp\{y_iQ(\theta_i)\},$$

where the parameter $Q(\theta_i)$ is the natural parameter.

Each subject $i$ has a $p \times 1$ vector of covariates, $X_i = (1, X_{i1}, \ldots, X_{ip-1})^T$. We
can link the parameter $Q(\theta_i)$ with a linear predictor $Q(\theta_i) = \eta_i = x_i^T \beta$, where $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$. This link function is called the canonical link because it expresses the mean of the response variable through the natural parameter.

A binomial distribution belongs to the exponential dispersion family. Suppose that $Y_i \sim Binomial(1, p_i)$. If we take

$$\log \frac{p_i}{1 - p_i} = \eta_i = x_i^T \beta,$$

this leads to a logistic regression model. It can be shown that the logit link function in the above model is actually the canonical link for the binomial responses. The statistical inferences for $glm$ are well established in the statistical literature. The standard software packages for $glm$ are widely available from most major software platforms such as $R$ and $SAS$. We will study the treatment-biomarker interaction effect for logistic regression models in chapter 4.

### 1.2.5 Penalized likelihood ratio test for mixture models

In generalized linear models, suppose we want to test the interaction between the treatment and the biomarker, in the case that the cut-point is an unknown parameter. The unknown cut-point of the biomarker divides the population into two mixture components (patient subsets) that respond differently to the new treatment. To analyze a typical mixture model, the penalized likelihood function is often used to overcome irregularities and eliminate non-identifiability encountered in the test.

Compared to the typical mixture model analysis, we deal with a mixture of regression models that is more challenging. In this section, we provide some useful materials corresponding to mixture models. Then, in chapter 4, we will propose a
new penalized likelihood method with a random penalty term to overcome irregularity and non-identifiability problems in the mixture distribution of two subpopulations described by regression models. Then, we conduct a test procedure for testing the biomarker main effect and the treatment-biomarker interaction effect.

Lindsay (1995) defined the simplest concept of the mixture models which arises when one samples from a population that consists of several homogeneous subpopulations. These subpopulations are called components of the population [49]. The number of components is generally denoted by $m$. In sampling from this population, $(X_i, J_i), i = 1, \ldots, n$, is the format of recorded data. Let $X_i = x_i$ be a measurement on the $i^{th}$ sampled unit, and $J_i = j$ indicates the index number of the component which the unit $i$ belongs to. The component density is defined as,

\[ P(X = x|J = j) = f(x, \theta, \xi_j), \]

where the variable $\xi_j$ is called the component parameter and is unknown. It describes the specific attributes of the $j^{th}$ component of the population. The vector $\theta$ is a vector of parameters that describes unknown parameters related to the entire population. The proportion of the total population that is in the $j^{th}$ component will be denoted by $\pi_j$, and is called the component weight. Therefore, $\sum_j \pi_j = 1$, and $\pi_j$'s are unknown parameters. Since we assume that the population has been sampled at random, the probability that an observation comes from the $j^{th}$ component is $\pi_j = P(J = j)$. We can describe $(X_i, J_i), i = 1, \ldots, n$ as a random sample from the following joint density,

\[ P(X = x, J = j) = P(X = x|J = j)P(J = j) = f(x, \theta, \xi_j)\pi_j. \]
1.2. STATISTICAL MODELS

The mixture model arises if the component label data is missing so that we only observe the sample \( x_1, \ldots, x_n \) from the marginal density of \( X \) (Lindsay (1989) [48]). Thus, the observed data set is a sample from the mixture density defined by,

\[
g(x, \pi, \theta) = \sum_{j=1}^{m} \pi_j f(x, \theta, \xi_j). \tag{1.7}
\]

In the context of mixture models, \( g \) is called a kernel function, and \( \pi = (\pi_1, \ldots, \pi_m) \) is a mixing distribution.

In studying mixture models, the ordinary likelihood ratio statistic often has complicated limiting distributions because of irregularities. The irregularity of mixture models has two reasons. First, the parameter space has boundaries, for example, \( 0 \leq \pi_j \leq 1 \), and second, the parameters are not fully identifiable. To overcome irregularities, the penalized likelihood ratio test is often used. The idea of the penalized likelihood ratio test is to penalize the likelihood function by adding a penalty term. The penalty term is defined such that there is no good fit when the parameter value gets close to the boundary. Chen, et. al. (2000) [14] defined the general format of the penalized log-likelihood as follows,

\[
p\ell_n(G) = \ell_n(G) - \lambda \psi(G),
\]

where \( G \) is mixing distribution function such as \( (\pi_1, \ldots, \pi_m) \) in (1.7), \( \ell_n(G) \) is the ordinary likelihood function, \( \lambda > 0 \) is a constant, and \( \psi(G) \) is a nonnegative function of \( G \). The penalty term \( \psi(G) \) usually becomes very large when the parameters get close to their boundaries.

The corresponding penalized log-likelihood ratio test statistic is \( 2p\ell_n(\hat{G}) - 2p\ell_n(\hat{G}_0) \),
where $\hat{G}, \hat{G}_0$ maximize \( p\ell_n(G) \) under the entire parameter space and the null parameter space, respectively. To carry out the statistical inference for the mixture model and parameters, we need to find the distribution of the penalized log-likelihood ratio test statistic (Chen (1998) [16]).

Using penalized likelihood ratio tests for mixture models has been extensively studied in the literature. Chen (1998) used the penalized likelihood ratio test for finite mixture models with multinomial observations [16]. Chen et.al. (2000) introduced a modified likelihood ratio test for homogeneity in the finite mixture models [14]. They used a penalized likelihood method to overcome irregularities in mixture models and obtained the asymptotic null distribution of the modified likelihood ratio test statistic. After that, the penalized likelihood ratio test was used to test for homogeneity in normal mixtures in the presence of a structural parameter (Chen and Chen (2003)[13]).

In chapter 4, we will propose a penalized likelihood ratio test for both biomarker main effect and treatment-biomarker interaction effect using a random penalty. Furthermore, by defining a new set of regularity conditions, we will study the properties and asymptotic distributions of the maximum penalized likelihood estimates. Moreover, using theoretical results and simulation studies, we will prove that the proposed test statistic has an asymptotic chi-square distribution, under the null hypothesis, that is, there is neither biomarker main effect nor treatment-biomarker interaction effect.
1.3 Tests for treatment-biomarker interaction

As we discussed in Section 1.1, testing for the treatment-biomarker interaction, with an unknown cut-point, is an important subject in biomarker study. To do the test, we need to develop a valid statistical test method. One of the most popular methods for statistical hypothesis testing is the Likelihood Ratio Test (LRT), because the likelihood ratio test statistic follows a Chi-square limiting distribution under regularity conditions (Serfling(1980), Wilks (1938) [60, 71]). However, the regularity conditions may not be satisfied when the cut-point of the biomarker is not known. For example, when the true parameters are located on the boundaries of the parameter space, the likelihood ratio test statistic results cannot be applied. Another example of irregularity occurs when we are using an indicator function in the model that divides the patients into two groups containing patients that benefit differently from the new treatment. In this case, the likelihood function is not a continuous and differentiable function with respect to the cut-point parameter. As a result, the maximum likelihood estimate of the cut-point value cannot be obtained easily, and the ordinary likelihood ratio test cannot be applied.

Therefore, it is very important to develop appropriate test statistics for testing the treatment-biomarker interaction effect in response to this need. As the distribution of the proposed test statistic is unknown, we propose two approaches to deal with this problem. First, we apply some resampling methods to approximate its distribution. Second, we develop a penalized likelihood ratio test statistic and obtain its asymptotic distribution by proving it theoretically.
1.4 Applications to clinical trials

1.4.1 Breast cancer data with Ki-67 as biomarker

The BIG 1-98 study is a Phase III clinical trial of 8,010 postmenopausal women with early stage invasive breast cancer. These patients were randomly assigned to one of four adjuvant therapy arms: Letrozole, Tamoxifen, or sequences of these agents. In this thesis, we are only interested in the patients who received either Letrozole or Tamoxifen treatments. The primary end-point of the trial was disease-free survival (DFS), which is defined as the length of time from randomization to the first event of invasive recurrence in local, regional, or distant sites, a new invasive breast cancer in the contralateral breast, any second nonbreast malignancy, or death as a result of any cause (Lazar et al. (2010) [45]).

A total of 4,922 patients were randomly assigned to receive 5 years of monotherapy with either Letrozole or Tamoxifen. The biomarker variable is Ki-67 labeling index, an indicator of tumor proliferation, that usually does not have a unit. A total of 2,685 patients had a tumor with material available for Ki-67 labeling index, with the following descriptive statistics. The mean value of Ki-67 index is 3.952, with median=3.580, $Q_1 = 2.980, Q_3 = 5.02$. The standard deviation of Ki-67 is equal to 1.448. The objective of this data analysis is to investigate the treatment effect (Letrozole vs Tamoxifen) on DFS with the presence of Ki-67 biomarker. From 2,685 patients available in the BIG 1-98 randomized clinical trial, $n_1 = 1,361$ of the patients received Letrozole, and $n_2 = 1,324$ of the patients received Tamoxifen.

Previous studies on the BIG 1-98 clinical trial show that Letrozole was more effective than Tamoxifen for patients with tumors expressing the highest levels of the Ki-67 labeling index (Viale et al. (2008), Lazar et al. (2010) [67, 45]). Moreover,
Lazar et al. (2010) [45] divided patients into two subgroups (to consider a binary biomarker) with high level of Ki-67 (> 10%) or low level of Ki-67 (≤ 10%). The used cut-point for Ki-67 biomarker labeling index is the median of the distribution of Ki-67 labeling index among all 4,922 patients from the BIG 1-98 clinical trial with available material. For patients with high tumor Ki-67, the hazards ratio of using Letrozole to Tamoxifen is $HR=0.53$; 95% C.I.=(0.39 to 0.72). Furthermore, among patients with low tumor Ki-67, the hazards ratio of using Letrozole to Tamoxifen is $HR=0.81$; with 95% C.I.=(0.57 ,1.15). That shows Letrozole is more effective than Tamoxifen for the patients with high values of Ki-67 labeling index (Viale et al. (2008) [67]). Lazar et al. (2010) [45] obtained p-value=0.11, indicating that the treatment-biomarker interaction is not statistically significant. Although the obtained p-value indicates that there is no statistically significant interaction, reduction in hazard ratios suggests the presence of interaction because Letrozole is acting differently from Tamoxifen for patients with higher Ki-67 labeling index.

In chapter 2, we apply the proposed residual bootstrap method to test the treatment-biomarker interaction for this clinical trial.

1.4.2 I-SPY 1 data with estrogen receptor total score (ER-TS) as biomarker

The I-SPY 1 trial (Investigation of Serial Studies to Predict Your Therapeutic Response With Imaging and Molecular Analysis) is a multicenter breast cancer study that integrates clinical, imaging, and genomic data to evaluate pathologic response and their relationship and predictability based on tumor biomarkers (Esserman et al. (2012) [27]). The main purpose was to determine whether response to therapy would predict Recurrence Free Survival (RFS). In general, pathologic Complete Response,
pCR (1 vs 0) can be used to predict disease free survival (DFS). The secondary purpose was developing a resource of clinical, molecular, genetic, and imaging biomarker data and a multicenter network to support high quality real-time biomarker evaluation for future trials (Esserman et al (2012) [27]).

Estrogen Receptor total score (ER-TS) is measured in an ordinal scale which ranges from 0 to 8, and most patients have ER-TS=0 (43%, n=72) or ER-TS=8 (35%, n=59).

Of the 237 patients enrolled, 221 patients from different institutes were available for analysis, 215 had pathology results after neoadjuvant therapy, and 210 had biomarker values available. Data were collected from 10 different institutions (University of North Carolina at Chapel Hill, University of Chicago, University of California at San Francisco, Memory Sloan-Kettering Cancer Center, Georgetown University, University of Texas South Western Medical center, University of Pennsylvania, University of Washington, University of Alabama at Birmingham, and Eastern Cooperative Oncology Group (ECOG)) with different numbers of patients from each institute. The data set is available from the National Cancer Informatics Program (NCIP website): https://cbiit.nci.nih.gov/ncip/biomedical-informatics-resources/data.

Esserman et al. (2012) [27] considered the association between pathology Complete Response (pCR) and recurrence free survival (RFS). For 172 patients who did not receive trastuzumab, the hazard ratios for RFS of patients with pCR vs. patients without pCR is HR=0.29, 95% C.I.=(0.07,0.82).

In chapter 3, we will investigate whether the associations between pCR and DFS are different among different subsets of patients defined by the ER total score (ER-TS) by taking the effect of centre clustering into account. To see that, we apply the
proposed cluster residual bootstrap method to test the pCR (treatment) and ER-TS (biomarker) interaction for clustered survival data in this clinical trial.

1.4.3 Prostate cancer data with acid phosphatase (AP) as biomarker

The prostate cancer clinical trial was conducted by the second Veterans Administration Cooperative Urologic Research Group (Byar and Corle (1977), Andrews and Herzberg (1985)[10, 5]). In this trial, \( n = 505 \) patients with prostate cancer were randomly assigned to receive either a placebo (\( n_0 = 128 \) patients) or diethylstilbestrol treatment (\( n_1 = 377 \) patients). The primary endpoint was the time from randomization to the time of death from any causes. Participants alive at the end of the study were censored at the last recorded time of being alive.

Byar and Corle (1977), Jiang et al. (2007) suggested that the serum prostatic acid phosphatase (AP) can be used as a predictive biomarker for diethylstilbestrol treatment, which is measured in King-Armstrong units ranging from 1 to 5960 [10, 39]. The log-rank test shows no significant difference in survival time distributions between the treatment and placebo groups (p-value=0.41). For the biomarker variable AP, the mean=101.6, median=7.00, while \( Q_1 = 5.00, Q_3 = 29.00 \). Also, the standard deviation is equal to 436.2071. Jiang et al. (2007) [39] applied an adaptive signature design to estimate the cut-point value for the AP biomarker. They obtained \( \hat{c} = 36 \), where \( c \) is an unknown cut-point for the AP biomarker, with a 95% C.I.=\((9,170)\). Moreover, Chen et al. (2014) [11] obtained \( \hat{c} = 46 \), with a 95% C.I.=\((27,107)\). This means that patients with AP > 46 benefit from the treatment, while patients with AP \( \leq 46 \) do not benefit from the treatment. Also, the treatment main effect is not significant, while the main effect of AP is highly significant. Both results in Jiang
et al. (2007) and Chen et al. (2014) [39, 11] are obtained in the absence of main
treatment effect.

In chapter 4, we apply the proposed penalized likelihood ratio test method to
data from this clinical trial in order to test for both the biomarker main effect and
the treatment-biomarker interaction effect in a logistic regression model, treating the
response variable as a binary variable.

1.5 Organization of thesis

Our goal in this research is to propose and to develop the bootstrap and the asymp-
totic methods to investigate if the new treatment benefits all patients in the same
way or not. We consider biomarker threshold models for both binary outcome and
survival outcome data when the biomarker variable is identified, but the cut-point
is unknown. In chapter 2, we will propose using a residual bootstrap method for
survival outcome data to test the treatment-biomarker interaction effect. In chap-
ter 3, the residual bootstrap test will be extended in a biomarker threshold model
to deal with clustered survival outcomes. In chapter 4, we will apply the penalized
likelihood ratio test for binary outcome data to overcome the irregularities caused by
boundary issues and the non-identifiability problem. Also, by proving some lemmas
and theorems, we will derive the asymptotic distribution of the penalized likelihood
ratio test statistic. Chapter 5 concludes and outlines future work.
Chapter 2

Residual bootstrap method for biomarker threshold models

2.1 Overview

In traditional clinical trials, randomized studies are often conducted to evaluate the treatment effects by including all eligible patients. However, in some subsets of patients with different characteristics, patients may respond differently to the treatments. A patient characteristic affecting a patient’s response to a certain treatment is called a predictive biomarker (Sargent, et al. (2005) [59]) or biomarker henceforth. The study of the interactive impacts of a biomarker on treatment outcomes is necessary for evaluating treatment effects within different biomarker defined patient subsets.

The conventional clinical trial design focuses on the overall treatment effect by including all eligible patients, which may fail to detect some stronger treatment effects restricted to subsets of patients. Clearly, identifying the subsets of patients who may not benefit, benefit less, or more from the new treatment can help to avoid unnecessary therapy and to make personalized decisions in treating patients. Biomarker threshold
models are frequently used (Simon and Maitournam (2004) [61]) to conduct this type of subset analysis. A biomarker is often measured on a continuous scale, such as Ki-67 for breast cancer patients (Lazar, et al. (2010) [45]). Royston et al. (2006) suggested that the biomarker should be treated as a continuous variable in order to use all potential information in the data [56]. Sargent et al. (2005) described a predictive (binary) biomarker that splits the patient population into two groups as either good or poor candidates for a specific treatment for optimistic treatment selection [59].

An important aspect of the biomarker study is to test for treatment-biomarker interaction effects. Existence of the interaction effects implies that the new treatment tends to have different effects on patients with different biomarker values, while no interaction effects implies that the new treatment has the same effects on all patients.

There are various methods in the literature for studying the biomarker-treatment interaction effects, for example, the traditional Cox regression model with the interaction, or the fractional polynomial approach for a continuous biomarker (Royston and Sauerbrei (2004) [57]). For a continuous biomarker, it is common in clinical trial analysis to assume that there is a biomarker cut-point that categorizes patients into subsets that benefit more, or do not benefit or benefit less from the new treatment. Jiang et al. (2007) proposed a biomarker adaptive threshold design for situations with a known biomarker, but an unknown cut-point defining two patient subsets with different treatment effects. They developed a test procedure based on the permutation method for detecting the difference in treatment effects between the two patient subsets [39]. In the same model framework, Chen et al. (2014) proposed a hierarchical Bayesian method for point and interval estimation of the biomarker cut-point and treatment effects on all patient subsets, but it cannot be used for testing
2.1. OVERVIEW


The classical likelihood ratio test for treatment-biomarker interaction effects cannot be applied for this kind of biomarker threshold model. This is because regularity conditions are required for the classical asymptotic results for the likelihood ratio test (Serfling (1980) [60]), while the biomarker threshold model is irregular in the presence of the unknown biomarker cut-point.

Upon a closer inspection of the permutation-based test for treatment-biomarker interaction effects by Jiang et al. (2007) [39], we notice that it is based on an implicit assumption that there are no main treatment effects in the model, in other words, the new treatment only benefits a subset of patients, and has no effects on the rest of the patients. This motivates us to develop a new testing method for the biomarker threshold model with an unknown biomarker cut-point, which is popular in biomarker-aided clinical trial studies. We propose a residual bootstrap method to approximate the distribution of a test statistic for identifying the treatment-biomarker interaction effects. We build the method based upon the bootstrap technique of Loughin (1995), while tackling the modeling challenges caused by the unknown biomarker cut-point [50]. The proposed method relaxes the model restriction of the permutation method by allowing main treatment effects, which is often necessary in common practice for modeling patient subsets that both benefit from the new treatment but at different levels. We then compare the proposed bootstrap method to the permutation method through extensive simulation studies.
2.2 Model and methods

2.2.1 Model

In clinical trials, patients are randomly assigned to either new treatment or control groups. Let $Z$ denote the treatment allocation variable, with $Z = 1$ for treatment and $Z = 0$ for the control groups. These two groups are compared with respect to time to a clinical event, such as death or disease progression. Let $\tilde{T}_i$ and $C_i$ be the potential failure time and censoring time for patient $i$, for $i = 1, \ldots, n$. Let $\delta_i = I(\tilde{T}_i < C_i)$ be a survival status indicator and $t_i = \min(\tilde{T}_i, C_i)$ be the observed failure or censoring time. Let $X$ be a biomarker variable. Without loss of generality, we assume that $0 \leq X \leq 1$. We further make an assumption that the biomarker is predictive of patient subsets with different treatment benefits. This can be described by the biomarker threshold model defined by Jiang, et. al. (2007) [39],

$$
\lambda_i(t) = \lambda_0(t) \exp \left\{ \beta_1 z_i + \beta_2 I(x_i > c) + \beta_3 z_i I(x_i > c) \right\}, \quad (2.1)
$$

where $z_i, x_i$ are observed values for $Z, X$ for patient $i$, $\lambda_0(t)$ is the baseline hazard function, and $\lambda_i(t)$ is the hazard function of patient $i$. In this model, $\beta_1$ is the main treatment effect, $\beta_2$ is the main biomarker effect, and $\beta_3$ is the treatment and biomarker interaction effect. The biomarker cut-point $c$ is also an unknown parameter and $0 \leq c \leq 1$. If $c$ is known, model (2.1) becomes a regular Cox model.

Specifically in model (2.1), on patients with biomarker values no greater than the cut-point $c$, the treatment effect in terms of log hazard ratio of treatment versus control groups is expressed by the parameter $\beta_1$; on patients with biomarker values greater than $c$, the treatment effect is expressed by $\beta_1 + \beta_3$. The treatment-biomarker
interaction effect $\beta_3$ is then the difference in treatment effects between the two patient subsets categorized by the biomarker and the cut-point. When $\beta_3 = 0$, all patients are expected to have the same treatment benefit ($\beta_1$) regardless of their biomarker values.

In this chapter, we are interested in studying the treatment-biomarker interaction effect in order to evaluate if patients in different biomarker subsets benefit differently from the new treatment. For this purpose, we focus on testing the existence of the treatment-biomarker interaction effect, with the null hypothesis $H_0 : \beta_3 = 0$ versus the alternative hypothesis $H_1 : \beta_3 \neq 0$. Define the null parameter space $\Theta_0 = \{ (\beta_1, \beta_2, \beta_3, c) : (\beta_1, \beta_2) \in \mathbb{R}^2, \beta_3 = 0, 0 < c < 1 \}$, and the entire parameter space $\Theta = \{ (\beta_1, \beta_2, \beta_3, c) : (\beta_1, \beta_2, \beta_3) \in \mathbb{R}^3, 0 < c < 1 \}$.

Let $Y_i(t) = I(t_i \geq t)$ be the indicator function that patient $i$ is at risk at time $t$. The partial likelihood function for model (2.1) for patients $i = 1, \ldots, n$ can be written as,

$$L(\beta, c) = \prod_{i=1}^{n} \left\{ \frac{\exp(x_i^T \beta)}{\sum_{k=1}^{n} Y_k(t_i) \exp(x_k^T \beta)} \right\}^{\delta_i}, \tag{2.2}$$

where $x_i^T = (z_i, I(x_i > c), z_i I(x_i > c))$, and $\beta = (\beta_1, \beta_2, \beta_3)^T$. Let $\ell$ be the logarithm of the partial likelihood function defined by (2.2). When the cut-point $c$ is unknown, the partial likelihood function is not continuous in $c$ and not differentiable with respect to $c$. As a consequence, it is difficult to estimate $c$ directly by maximizing the partial likelihood function, and the classical asymptotic results for the likelihood ratio test statistic are not valid.

For a given value of $c$, denote the corresponding log likelihood ratio statistic by

$$LR(c) = 2 \left\{ \ell(\hat{\beta}_1c, \hat{\beta}_2c, \hat{\beta}_3c; c) - \ell(\hat{\beta}_1c, \hat{\beta}_2c, 0; c) \right\},$$
2.2. MODEL AND METHODS

where \( \hat{\beta}_1c, \hat{\beta}_2c, \) and \( \hat{\beta}_3c \) are the maximum partial likelihood estimates obtained under the entire parameter space for a given \( c \), \( \Theta_c = \{ (\beta_1, \beta_2, \beta_3) \in \mathbb{R}^3 \} \), and \( \tilde{\beta}_1c, \tilde{\beta}_2c \) are the maximum partial likelihood estimates under the null parameter space \( \Theta_0c = \{ (\beta_1, \beta_2) \in \mathbb{R}^2, \beta_3 = 0 \} \) for a given \( c \). We then define a test statistic for testing \( H_0 \) in the form of

\[
LR = \max_{0 < c < 1} LR(c),
\]

which is the maximum of the log likelihood ratio statistic \( LR(c) \) over any possible \( c \) value, \( 0 < c < 1 \).

The distribution of the proposed test statistic is unknown in this setting. To conduct statistical inferences for the test, we use resampling methods to approximate the distribution of \( LR \), under the null hypothesis \( H_0 : \beta_3 = 0 \). For this purpose, we propose and study the residual bootstrap method in Section 2.2.2. To compare the proposed method and the existing permutation method, we also briefly review an existing method based on the permutation test (Jiang, et al. (2007) [39]) in Section 2.2.3.

2.2.2 Residual bootstrap method

Loughin (1995) proposed a residual bootstrap method for the Cox proportional hazards regression model when explanatory variables are nonrandom constants fixed by the design [50]. We extend the method to the biomarker threshold model (2.1) in the presence of an unknown cut-point, which is beyond the scope of traditional regression models for survival data. The test for the treatment-biomarker interaction effect in model (2.1), relies on a complicated and unconventional test statistic in (2.3), which needs to be handled carefully in the bootstrap procedure with new strategies.
The proposed sampling scheme by Loughin (1995) [50] is based on the fact that, when the relative risk function is independent of time, the Cox’s partial likelihood function defined by (2.2) remains invariant under monotone increasing transformations of time $t$. For a Cox’s model, we can write,

$$S(t) = S_0(t)^{\exp(x^T\beta)} = \{1 - F_0(t)\}^{\exp(x^T\beta)},$$

where $S_0(t)$ is the baseline survival function, and $F_0(t) = 1 - S_0(t)$ is the baseline cumulative distribution function. Therefore,

$$F_0(t) = 1 - \{S(t)\}^{\exp(-x^T\beta)}$$

is a monotone increasing function of $t$. The invariance property of the partial likelihood function implies replacing failure time $t$ by a monotone transformation $F_0(t)$ in the data will not change the statistical inference.

Direct application of the method by Loughin (1995) [50] is impossible for the biomarker threshold model (2.1) because of unknown cut-point $c$. Instead, we propose the following residual bootstrap test (RBT) method with four steps for testing the treatment-biomarker interaction effect with $H_0 : \beta_3 = 0$, by meticulously incorporating the profile likelihood estimation for $c$.

**Step 1.** Estimate $c$ by considering a grid of $c$ values, $0 < c < 1$. Obtain the maximum profile likelihood estimate of $c$,

$$\hat{c} = \arg \max_{0 < c < 1} \ell(\hat{\beta}_1; \hat{\beta}_2; \hat{\beta}_3; c),$$
2.2. MODEL AND METHODS

where \( \hat{\beta}_1c, \hat{\beta}_2c, \) and \( \hat{\beta}_3c \) are the maximum partial likelihood estimates of \( \beta_1, \beta_2, \) and \( \beta_3 \) for each given value of \( c \), under the entire parameter space \( \Theta_c \). Then, taking \( c = \hat{c} \), fit model (2.1) as a Cox model based on the original data \((t_i, \delta_i, x_i), i = 1, \ldots, n\), to obtain \( \hat{\beta}_1, \hat{\beta}_2, \) and \( \hat{\beta}_3 \). Estimate the survival probabilities \( \hat{S}_i(t_i) = \hat{S}(t_i), i = 1, \ldots, n \), using the Nelson-Aalen Method [1, 53].

Step 2. Generate bootstrap data \((\hat{u}^*_i, \delta^*_i, x_i)\) by sampling with replacement from \((\hat{u}_i, \delta_i, x_i), i = 1, \ldots, n\). Calculate the probability-scale failure time, under \( H_0 \),

\[ y^*_i = 1 - (\hat{u}^*_i)^{\exp[-(z_i \hat{\beta}_1 + I(x_i > \hat{c}) \hat{\beta}_2)]}, \]

where \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) are the maximum partial likelihood estimates, and \( \hat{c} \) is the maximum profile likelihood estimate of \( c \), under the null parameter space \( \Theta_0 \). In light of (2.4), \( y^* \) is a resampled version of \( F_0(t) \).

Step 3. For each given value of \( c \), fit two models on the entire parameter space \( \Theta_c \) and the null parameter space \( \Theta_0c \) for the bootstrap data \((y^*_i, \delta^*_i, x_i), i = 1, \ldots, n\), \((y^* \) as survival time and \( \delta^* \) as censoring indicator) to calculate the likelihood ratio test statistic \( LR^*_b(c) \). Across a grid of \( c \) values, \( 0 < c < 1 \), obtain the bootstrap version of the test statistic for the \( b \)th bootstrap sample,

\[ LR^*_b = \max_{0 < c < 1} LR^*_b(c). \]

Step 4. Repeat Steps 2, 3 \( B \) times, and obtain \( B \) replications of \( LR^*_b, b = 1, \ldots, B \).

The empirical distribution of \( \{LR^*_b\}_{b=1}^B \) provides an approximation to the null distribution of the test statistic \( LR \) given by (2.3). The \( p \)-value of the RBT is the
2.2. MODEL AND METHODS

proportion of $LR^*$ values greater than the observed $LR$, 

$$p\text{-value}_{RBT} = \frac{1}{B} \sum_{b=1}^{B} I(LR^*_b > LR), \quad (2.5)$$

where $LR$ is the observed test statistic of the form (2.3) calculated based on the original data.

2.2.3 Permutation test

Jiang et al. (2007) introduced an adaptive design to investigate if the new treatment benefits all patients or only a subset of patients (treatment-biomarker interaction) [39]. In this subsection, we review the permutation test (PT) method of Jiang et al. (2007) [39] as another way to approximate the distribution of the proposed test statistic $LR$ in (2.3) for testing the treatment-biomarker interaction effect.

The PT method consists of $B$ permutation runs. In each run $p$, a permuted version of data is created by randomly assigning patients to the new and control treatments, while keeping other variables the same. On the permuted data of $p^{th}$ run, fit two Cox models under the entire parameter space and the null parameter space for a grid of $c$ values, calculate the likelihood ratio test statistic $LR^p_b(c)$, and obtain the test statistic,

$$LR^p_b = \max_{0<c<1} LR^p_b(c).$$

With $B$ permutation replications, $p = 1, \ldots, B$, obtain the $p$-value 

$$p\text{-value}_{PT} = \frac{1}{B} \sum_{b=1}^{B} I(LR^p_b > LR), \quad (2.6)$$
where $LR$ of the form in (2.3) is obtained based on the original data.

We notice that the PT method for model (2.1) randomly assigns patients to treatments, hence implicitly assumes that observations are exchangeable under the null hypothesis, i.e., there is no treatment main effect. In other words, the PT method is considered to be correct only when $\beta_1 = \beta_3 = 0$. Jiang et al. (2007) has not addressed this restriction, and has not examined the PT method when the assumption of $\beta_1 = 0$ is violated [39]. The proposed RBT method has no such restrictions.

\section*{2.3 Simulation studies}

\subsection*{2.3.1 Simulation model design}

We conducted simulation studies to evaluate the finite sample performance of the RBT method for testing $H_0 : \beta_3 = 0$ in model (2.1) in comparison to the PT method. We generated survival time from a Weibull distribution with hazard function

$$
\lambda_i(t) = \nu \gamma (\gamma t)^{\nu-1} \exp \left\{ \beta_1 z_i + \beta_2 I(x_i > c) + \beta_3 z_i I(x_i > c) \right\},
$$

(2.7)

with parameters $\gamma, \nu$, and baseline hazard function $\lambda_0(t) = \nu \gamma (\gamma t)^{\nu-1}$. The biomarker values $x_i$'s were generated from a Uniform distribution on the interval (0,1). Here, we used a censoring time generated from a Uniform distribution on the interval (0,1.5), which results in a censoring rate between 3%-75%, depending on the setting of the parameters.

We considered three different designs for assigning patients to treatments in order to compare the performance of the RBT and PT methods in different situations. In design I, we applied an unequal randomization which assigns 80% of the patients to the
treatment arm and 20% to the control arm. This type of design is particularly useful when existing evidence suggests that the new treatment may have other additional clinical benefit, for example, less toxicity compared with the standard treatment. In design II, we assumed that patients can be divided into two groups: high risk and low risk groups before randomization (Barker et al. (2009), Zang and Lee (2015) [6, 72]). We assumed 50% of patients are in the high risk group and the other 50% in the low risk group, respectively. For the high risk group patient, the new treatment was assigned to the patient with probability 0.75. Otherwise, for the low risk group, the new treatment was assigned to the patient with probability 0.5. Design II is based on the consideration that, for the high risk group, the treatment is supposed to benefit this group of patients more. Therefore, it is reasonable to have a bigger percentage of high risk patients to receive the new treatment than the standard treatment or control. In design III, we considered a typical randomization and assigned the two treatments to the patients with equal weights.

To generate the simulated data set, we considered different combinations of $\beta_1$, $\beta_2$, and $\beta_3 = \log(0.1), \log(0.3), \log(0.5), \log(0.9)$, and the true cut-point values $c_0 = 0.25, 0.6$, and 0.75. We let $\nu = 1.5, 2.0$, and $\gamma = 2.0$ in generating the Weibull failure time. To calculate the test statistic in (2.3), we used a grid search of candidate biomarker values. We repeated the simulation $R = 500$ times for each parameter combination.

2.3.2 Simulation results

For both the bootstrap and permutation methods, we used $B = 200$ bootstrap/permutation replications, respectively. The results for empirical test sizes for both RBT and PT
methods for designs I-III are presented in Table 2.1. The empirical test size is the percentage of the simulation replications that reach the rejection criterion, i.e. reach the pre-specified level of the statistical significance ($\alpha = 0.05$) when data are generated under the null hypothesis, $H_0$. To calculate the empirical test sizes for RBT and PT methods, we calculated p-values from (2.5) and (2.6) such that the test size is the proportion of p-values that are smaller than $\alpha$. If the empirical size of the test turns out to be close to the significance level ($\alpha = 0.05$), we say that the proposed test procedure can control the type I error.
2.3. SIMULATION STUDIES

Table 2.1: The empirical size of the test, $H_0 : \beta_3 = 0^1$.

<table>
<thead>
<tr>
<th>$\nu$</th>
<th>$c_0$</th>
<th>$\exp(\beta_1)$</th>
<th>$\exp(\beta_2)$</th>
<th>Design I</th>
<th>Design II</th>
<th>Design III</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PT RBT</td>
<td>PT RBT</td>
<td>PT RBT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.25</td>
<td>0.1 0.5</td>
<td>0.064 0.050</td>
<td>0.060 0.054</td>
<td>0.062 0.056</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.25</td>
<td>0.5 0.3</td>
<td>0.048 0.052</td>
<td>0.060 0.046</td>
<td>0.034 0.022</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.25</td>
<td>0.9 0.3</td>
<td>0.048 0.038</td>
<td>0.058 0.048</td>
<td>0.046 0.028</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.25</td>
<td>0.5 0.5</td>
<td>0.038 0.038</td>
<td>0.052 0.046</td>
<td>0.044 0.032</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.60</td>
<td>0.1 0.5</td>
<td>0.056 0.064</td>
<td>0.054 0.048</td>
<td>0.080 0.082</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.60</td>
<td>0.5 0.3</td>
<td>0.052 0.052</td>
<td>0.088 0.064</td>
<td>0.040 0.044</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.60</td>
<td>0.9 0.3</td>
<td>0.058 0.070</td>
<td>0.114 0.068</td>
<td>0.042 0.052</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.75</td>
<td>0.5 0.3</td>
<td>0.044 0.050</td>
<td>0.058 0.060</td>
<td>0.048 0.052</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.75</td>
<td>0.9 0.3</td>
<td>0.046 0.056</td>
<td>0.068 0.062</td>
<td>0.046 0.022</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.75</td>
<td>0.5 0.5</td>
<td>0.048 0.050</td>
<td>0.054 0.046</td>
<td>0.046 0.050</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.25</td>
<td>0.1 0.5</td>
<td>0.064 0.056</td>
<td>0.068 0.062</td>
<td>0.044 0.036</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.25</td>
<td>0.3 0.5</td>
<td>0.046 0.040</td>
<td>0.052 0.042</td>
<td>0.048 0.048</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.25</td>
<td>0.5 0.3</td>
<td>0.054 0.054</td>
<td>0.048 0.042</td>
<td>0.052 0.038</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.25</td>
<td>0.9 0.3</td>
<td>0.048 0.038</td>
<td>0.060 0.048</td>
<td>0.042 0.040</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.25</td>
<td>0.5 0.5</td>
<td>0.048 0.048</td>
<td>0.054 0.036</td>
<td>0.048 0.036</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.60</td>
<td>0.1 0.5</td>
<td>0.042 0.038</td>
<td>0.072 0.060</td>
<td>0.044 0.058</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.60</td>
<td>0.5 0.3</td>
<td>0.070 0.058</td>
<td>0.086 0.044</td>
<td>0.050 0.056</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.75</td>
<td>0.5 0.3</td>
<td>0.058 0.058</td>
<td>0.048 0.048</td>
<td>0.056 0.052</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.75</td>
<td>0.9 0.3</td>
<td>0.048 0.058</td>
<td>0.060 0.050</td>
<td>0.044 0.054</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.75</td>
<td>0.5 0.5</td>
<td>0.048 0.052</td>
<td>0.052 0.038</td>
<td>0.046 0.054</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.75</td>
<td>0.9 0.3</td>
<td>0.048 0.048</td>
<td>0.054 0.044</td>
<td>0.032 0.030</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0.25</td>
<td>0.1 0.5</td>
<td>0.060 0.044</td>
<td>0.066 0.062</td>
<td>0.032 0.072</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0.25</td>
<td>0.5 0.3</td>
<td>0.056 0.050</td>
<td>0.060 0.036</td>
<td>0.072 0.054</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0.25</td>
<td>0.9 0.3</td>
<td>0.040 0.036</td>
<td>0.052 0.038</td>
<td>0.042 0.036</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0.25</td>
<td>0.5 0.5</td>
<td>0.046 0.048</td>
<td>0.050 0.042</td>
<td>0.044 0.032</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0.60</td>
<td>0.1 0.5</td>
<td>0.040 0.028</td>
<td>0.070 0.048</td>
<td>0.042 0.048</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0.60</td>
<td>0.5 0.3</td>
<td>0.058 0.056</td>
<td>0.080 0.030</td>
<td>0.048 0.048</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0.60</td>
<td>0.9 0.3</td>
<td>0.044 0.050</td>
<td>0.130 0.046</td>
<td>0.046 0.040</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0.75</td>
<td>0.5 0.3</td>
<td>0.060 0.054</td>
<td>0.058 0.046</td>
<td>0.052 0.056</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0.75</td>
<td>0.9 0.3</td>
<td>0.050 0.064</td>
<td>0.054 0.038</td>
<td>0.046 0.042</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0.75</td>
<td>0.5 0.5</td>
<td>0.052 0.048</td>
<td>0.052 0.052</td>
<td>0.048 0.046</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.25</td>
<td>0.1 0.5</td>
<td>0.058 0.034</td>
<td>0.070 0.044</td>
<td>0.060 0.056</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.25</td>
<td>0.3 0.5</td>
<td>0.054 0.044</td>
<td>0.042 0.044</td>
<td>0.042 0.036</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.25</td>
<td>0.5 0.3</td>
<td>0.058 0.046</td>
<td>0.054 0.044</td>
<td>0.052 0.038</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.25</td>
<td>0.9 0.3</td>
<td>0.042 0.042</td>
<td>0.056 0.038</td>
<td>0.040 0.038</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.25</td>
<td>0.5 0.5</td>
<td>0.056 0.050</td>
<td>0.032 0.034</td>
<td>0.050 0.028</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.60</td>
<td>0.1 0.5</td>
<td>0.046 0.026</td>
<td>0.062 0.038</td>
<td>0.060 0.056</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.60</td>
<td>0.5 0.3</td>
<td>0.046 0.036</td>
<td>0.102 0.032</td>
<td>0.038 0.032</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.75</td>
<td>0.5 0.3</td>
<td>0.052 0.046</td>
<td>0.060 0.050</td>
<td>0.046 0.044</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.75</td>
<td>0.9 0.3</td>
<td>0.052 0.058</td>
<td>0.064 0.044</td>
<td>0.060 0.062</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.75</td>
<td>0.5 0.5</td>
<td>0.054 0.048</td>
<td>0.040 0.046</td>
<td>0.038 0.048</td>
<td></td>
</tr>
</tbody>
</table>

\[1 \text{ Significance level } \alpha = 0.05. \text{ Failure time follows Weibull distribution in (2.7) with } \gamma = 2.0, n = 500, B = 200 \text{ bootstrap samples, } R = 500 \text{ replications, under the null hypothesis.} \]
Simulation results in Table 2.1 indicate that the RBT method provides correct test sizes, while the PT method cannot control the type I error in some cases, especially for a complicated design such as design II. For some combinations of parameter settings, the test based on PT method gives wrong test size. For example, when \( \nu = 2.0, c_0 = 0.6, \beta_1 = \log(0.5), \beta_2 = \log(0.3) \), using design II, the empirical test size for PT is 0.102, while the empirical test size for RBT is 0.032. This is not surprising because the PT method requires the model assumption of no main treatment effect (\( \beta_1 = 0 \)), which apparently does not hold here. The PT method proposed by Jiang et al. (2007) should not be applied unless there is clear prior knowledge that no main treatment effects exist [39]. The RBT method is a valid approach without this model restriction.
2.3. SIMULATION STUDIES

Table 2.2: The empirical power of RBT, $H_1: \beta_3 \neq 0^1$.

<table>
<thead>
<tr>
<th>$\nu$</th>
<th>$c_0$</th>
<th>$\exp(\beta_1)$</th>
<th>$\exp(\beta_2)$</th>
<th>$\exp(\beta_3)$</th>
<th>Design I</th>
<th>Design II</th>
<th>Design III</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.578</td>
<td>0.552</td>
<td>0.544</td>
</tr>
<tr>
<td>0.5</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
<td>0.404</td>
<td>0.404</td>
<td>0.436</td>
</tr>
<tr>
<td>0.5</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.528</td>
<td>0.498</td>
<td>0.596</td>
</tr>
<tr>
<td>0.5</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.478</td>
<td>0.414</td>
<td>0.572</td>
</tr>
<tr>
<td>0.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.610</td>
<td>0.666</td>
<td>0.740</td>
</tr>
<tr>
<td>0.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.392</td>
<td>0.436</td>
<td>0.536</td>
</tr>
<tr>
<td>1.0</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.748</td>
<td>0.790</td>
<td>0.792</td>
</tr>
<tr>
<td>1.0</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
<td>0.542</td>
<td>0.596</td>
<td>0.846</td>
</tr>
<tr>
<td>1.0</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.706</td>
<td>0.682</td>
<td>0.790</td>
</tr>
<tr>
<td>1.0</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.664</td>
<td>0.600</td>
<td>0.776</td>
</tr>
<tr>
<td>1.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.820</td>
<td>0.854</td>
<td>0.926</td>
</tr>
<tr>
<td>1.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.620</td>
<td>0.833</td>
<td>0.774</td>
</tr>
<tr>
<td>1.5</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.784</td>
<td>0.874</td>
<td>0.902</td>
</tr>
<tr>
<td>1.5</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
<td>0.588</td>
<td>0.680</td>
<td>0.710</td>
</tr>
<tr>
<td>1.5</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.792</td>
<td>0.806</td>
<td>0.914</td>
</tr>
<tr>
<td>1.5</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.808</td>
<td>0.748</td>
<td>0.916</td>
</tr>
<tr>
<td>1.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.936</td>
<td>0.970</td>
<td>0.994</td>
</tr>
<tr>
<td>1.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.776</td>
<td>0.838</td>
<td>0.924</td>
</tr>
<tr>
<td>2.0</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.796</td>
<td>0.912</td>
<td>0.930</td>
</tr>
<tr>
<td>2.0</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
<td>0.554</td>
<td>0.696</td>
<td>0.692</td>
</tr>
<tr>
<td>2.0</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.856</td>
<td>0.858</td>
<td>0.964</td>
</tr>
<tr>
<td>2.0</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.876</td>
<td>0.824</td>
<td>0.966</td>
</tr>
<tr>
<td>2.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.986</td>
<td>0.996</td>
<td>0.998</td>
</tr>
<tr>
<td>2.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.808</td>
<td>0.908</td>
<td>0.974</td>
</tr>
</tbody>
</table>

$^1$ Significance level $\alpha = 0.05$. Failure time follows Weibull distribution in (2.7) with $\gamma = 2.0$, $n = 500$, $B = 200$ bootstrap samples, $R = 500$ replications, under the alternative hypothesis.

Furthermore, we evaluated the empirical power of the RBT method under the alternative hypothesis for different values of $\beta_3 \neq 0$, and the results are presented in Table 2.2. The empirical power of the test is the percentage of the simulated replications that reach the pre-specified significance level ($\alpha = 0.05$), when data are generated under the alternative hypothesis. Since the PT method has restrictive model assumption and cannot control type I error in all scenarios, it is meaningless to further examine its power. The empirical power of RBT, as seen in Table 2.2, is quite satisfactory.
when testing for the treatment-biomarker interaction effects in various situations.

Table 2.3: Empirical bias and standard error of the profile likelihood estimate of $c^1$.

<table>
<thead>
<tr>
<th>$\nu$</th>
<th>$c_0$</th>
<th>$e^{\beta_1}$</th>
<th>$e^{\beta_2}$</th>
<th>$e^{\beta_3}$</th>
<th>Bias($\hat{c}$)</th>
<th>S.E.($\hat{c}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.0032</td>
<td>0.026</td>
</tr>
<tr>
<td>0.5</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.0016</td>
<td>0.013</td>
</tr>
<tr>
<td>0.5</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>-0.00046</td>
<td>0.022</td>
</tr>
<tr>
<td>0.5</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>-0.0002</td>
<td>0.008</td>
</tr>
<tr>
<td>0.5</td>
<td>0.75</td>
<td>0.3</td>
<td>0.5</td>
<td>0.3</td>
<td>-0.0004</td>
<td>0.021</td>
</tr>
<tr>
<td>0.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.0006</td>
<td>0.013</td>
</tr>
<tr>
<td>1.0</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.0012</td>
<td>0.018</td>
</tr>
<tr>
<td>1.0</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.0004</td>
<td>0.012</td>
</tr>
<tr>
<td>1.0</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>-0.0003</td>
<td>0.011</td>
</tr>
<tr>
<td>1.0</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>-0.0004</td>
<td>0.005</td>
</tr>
<tr>
<td>1.0</td>
<td>0.75</td>
<td>0.3</td>
<td>0.5</td>
<td>0.3</td>
<td>-0.0003</td>
<td>0.015</td>
</tr>
<tr>
<td>1.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.0008</td>
<td>0.012</td>
</tr>
<tr>
<td>1.5</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.0008</td>
<td>0.014</td>
</tr>
<tr>
<td>1.5</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>-0.0012</td>
<td>0.010</td>
</tr>
<tr>
<td>1.5</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>-0.0004</td>
<td>0.008</td>
</tr>
<tr>
<td>1.5</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>-0.0002</td>
<td>0.004</td>
</tr>
<tr>
<td>1.5</td>
<td>0.75</td>
<td>0.3</td>
<td>0.5</td>
<td>0.3</td>
<td>-0.0008</td>
<td>0.013</td>
</tr>
<tr>
<td>1.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>-0.0026</td>
<td>0.010</td>
</tr>
<tr>
<td>2.0</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>-0.0014</td>
<td>0.012</td>
</tr>
<tr>
<td>2.0</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>-0.0034</td>
<td>0.009</td>
</tr>
<tr>
<td>2.0</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.00004</td>
<td>0.006</td>
</tr>
<tr>
<td>2.0</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>-0.0001</td>
<td>0.003</td>
</tr>
<tr>
<td>2.0</td>
<td>0.75</td>
<td>0.3</td>
<td>0.5</td>
<td>0.3</td>
<td>-0.0025</td>
<td>0.010</td>
</tr>
<tr>
<td>2.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>-0.0028</td>
<td>0.010</td>
</tr>
</tbody>
</table>

Superscript 1: Failure time were generated from Weibull distribution in (2.7) with $\gamma = 2.0$, under $H_1: \beta_3 \neq 0$, for design I, with $n = 500$, $B = 200$ bootstrap samples, $R = 500$ replications.

In addition, we also explored the finite sample properties for the profile partial likelihood estimate of $c$. To find the profile estimate of $c$, we used a grid search of $c$ values on the interval $(0.1, 0.9)$ increased by 0.01. Table 2.3 shows the empirical bias and standard error of the profile partial likelihood estimate of $c$ for different
combinations of the parameters and design I, under the alternative hypothesis. For this table, Weibull failure times were generated from model (2.7) with $\nu = 1.5, 2.0$ and $\gamma = 2.0$ in the baseline hazard function. The profile partial likelihood method turns out to be quite accurate for estimating $c$, with very small bias and standard error.

As discussed in Section 2.1 and Section 2.2, the classical likelihood ratio test is not valid for testing the treatment-biomarker interaction effects in model (2.1) in the presence of an unknown biomarker cut-point. As an important illustration, we simulated data sets based on model (2.1) with $c = 0.25$, $\beta_1 = \log(0.3)$, $\beta_2 = \log(0.5)$, $\beta_3 = 0$, and directly applied the classical likelihood ratio test for $H_0 : \beta_3 = 0$; the resulting test size was 0.264, which is obviously wrong for the significance level $\alpha = 0.05$.

### 2.4 Application to breast international group (BIG) 1-98 data set

To demonstrate how to apply the RBT method for testing the interaction between the biomarker and the treatment, we applied the method to a data set from Breast International Group (BIG) 1-98 randomized clinical trial (Lazar et al. (2010) [45]). This dataset is available from the R software package “stepp”. The BIG 1-98 is a Phase III clinical trial of 8010 postmenopausal women with hormone-receptor-positive early invasive breast cancer who were randomly assigned adjuvant therapy of Letrozole as new treatment or Tamoxifen as control treatment. Among the 8010 patients, 2685 of them have available Ki-67 biomarker measurements.

The Ki-67 biomarker is a labeling index measurement of cell proliferation (Note that an index usually does not have a unit). Previous study shows that Letrozole
was more effective than Tamoxifen for patients with tumors expressing the highest levels of the Ki-67 labeling index. Lazar et al. (2010) obtained that the $p$-value=$0.09$ (Fig. 1(c) of Lazar et al. (2010) [45]) for the interaction term in hazard rate based on the Subpopulation Treatment Effect Pattern Plot (STEPP) method introduced by Bonetti and Gelber (2000)[9].

To apply the proposed RBT method, we first transformed the Ki-67 biomarker value to the interval $(0, 1)$ using empirical percentiles of the original measurement. Then, we applied the RBT method to the data with $B = 2000$ replications, which gives a $p$-value of $0.0505$ for the treatment and biomarker interaction effect. For significance level $\alpha = 0.1$, this confirms the results of comparing hazard ratios by Lazar et al. (2010) [45] that there is a significant interaction effect between the treatment and the Ki-67 biomarker such that for patients with high Ki-67 labeling index, Letrozole is more effective than Tamoxifen.
Figure 2.1: Histogram of the test statistic values based on the residual bootstrap test for data from the Breast International Group BIG 1-98 clinical trial, where $LR^*$ and $LR$ are the test statistic of the form (2.3) based on bootstrapped and original data, respectively.
2.4. APPLICATION TO BREAST INTERNATIONAL GROUP (BIG) 1-98 DATA SET

Figure 2.1 shows the histogram of the bootstrap test statistic samples under the null hypothesis based on the RBT method, which provides an empirical distribution of the proposed likelihood ratio test statistic $LR$. As displayed in the graph, the p-value of RBT is the proportion of resampled data with $LR^*$ greater than the observed $LR$.

We used a grid search of $c$ values between $(0.01, 1)$ increased by 0.01 to obtain the profile partial likelihood estimate of the biomarker cut-point. The profile partial likelihood estimate of the cut-point value is 0.07 in the transformed scale, which is equivalent to $\hat{c} = 2.37$ in the original scale. From Figure 2.2, we can infer that patients with original Ki-67 values greater than 2.37 may benefit more from the Letrozole treatment. Furthermore, Figure 2.2 illustrates the profile partial likelihood function $\ell(\hat{\beta}_c, c)$ versus transformed threshold parameter $c$ (left panel) and original values of Ki-67 biomarker (right panel), for data from Breast International Group BIG 1-98 clinical trial.
2.4. APPLICATION TO BREAST INTERNATIONAL GROUP (BIG)
1-98 DATA SET

Figure 2.2: Profile partial likelihood $\ell(\hat{\beta}_c, c)$ versus transformed threshold parameter $c$ (left panel) and original values of Ki-67 (right panel), under $H_1$, for data from the Breast International Group BIG 1-98 clinical trial.
We further calculate hazard ratio (HR) of Letrozole versus Tamoxifen for both subsets defined by Ki-67. For patients with Ki-67 > 2.37, HR=0.6993, 95% confidence interval=(0.512,0.955), p-value=0.0246. For patients with Ki-67 ≤ 2.37, HR=1.387, 95% confidence interval=(0.975,1.973), p-value=0.069.

![KM Curves with Ki-67>2.37](image1)

![KM curves with Ki-67<=2.37](image2)

Figure 2.3: Kaplan-Meier estimates of disease-free survival for subsets defined by the biomarker Ki-67 using $\hat{c} = 2.37$.

Moreover, Figure 2.3 presents the Kaplan-Meier estimates of disease-free survival for subsets defined by Ki-67 using $\hat{c} = 2.37$. Considering both hazard ratio and Figure 2.3, we can say patients receiving Letrozole in subset with Ki-67 > 2.37 will have lower hazards, i.e., Letrozole is more effective than Tamoxifen for patients with biomarker values greater than 2.37.
2.5 Discussions

In clinical trials, testing for treatment-biomarker interaction effects is important when a new treatment tends to benefit a subset of patients more. In this chapter, we considered the popular biomarker threshold model (2.1), which models the biomarker as a binary variable with an unknown cut-point, and in this way categorizes patients into two subsets with different treatment effect levels. The classical likelihood ratio test does not work for testing the treatment-biomarker interaction effects in model (2.1), because the unknown biomarker cut-point causes model irregularity and discontinuous likelihood function. The PT method by Jiang et al. (2007) [39] has an implicit model restriction and is only appropriate when no main treatment effects exist ($\beta_1 = 0$). We proposed the RBT method to test for the treatment-biomarker interaction effects. It extends the residual bootstrap for Cox proportional hazards regression model proposed by Loughin (1995) [50] to the biomarker threshold model (2.1), which is more complex than the conventional Cox model framework. In the RBT method, we applied the profile likelihood method in the estimation of the biomarker cut-point $c$. Although the profile partial likelihood method provides a good estimate (with small biasedness) for the cut-point parameter $c$ (Chen et al. (2014) and Pons (2003) [11, 54]), it cannot be used directly for testing the treatment-biomarker interaction effects. In order to deal with the unknown biomarker cut-point, we built the RBT method by incorporating the profile partial likelihood estimation of $c$ in the bootstrap layers and in the inference procedure whenever necessary.

We considered the simulation scenarios with the existence of main treatment effects ($\beta_1 \neq 0$), so that the model assumption for the permutation test method is violated. The empirical test sizes of the PT method are sometimes reasonable, but
sometimes far off the nominal significance level. The PT method is not generally applicable because of its model restrictions. The proposed RBT method gives correct test size without such restrictions, and it provides good power for detecting departure from the null hypothesis. It adds a new and general tool to clinical trial studies with survival outcomes to identify varying treatment effects utilizing patient’s biomarker information.
Chapter 3

Residual bootstrap method for clustered survival outcome

3.1 Overview

In modern cancer clinical trials, a large sample size is often required to detect small but clinically meaningful effects. As a result, many clinical trials often involve patients from multiple institutions. In this situation, patients may belong to different clusters (family, hospital, clinical center, etc.) such that the patients in a certain cluster are dependent due to the shared environmental and genetic factors. In this chapter, we will develop residual bootstrap methods to test the treatment-biomarker interaction for clustered survival outcome.

In survival analysis, frailty models based on the likelihood function and marginal models based on generalized estimation equations (GEE) are two (Kalbfleisch and Prentice (2002) [42]) main approaches for analyzing clustered data. Liang (1987) proposed mixed-effect models to study samples grouped in m potential clusters [47]. A specific outcome variable $Y_{ij}$ for individual $j$ in cluster $i$, containing a group of samples, has a density function $g_{ij}(y_{ij}, \beta, w_i)$. In such a model, the parameter $\beta$
represents the fixed effects of covariates, while $\boldsymbol{w} = (w_1, \ldots, w_m)^T$ represents the vector of random effects related to $m$ clusters.

There is an extensive literature on modeling clustered survival outcome data through both parametric and semi-parametric frailty models. While patients from the same cluster cannot be treated as independent samples, Glidden and Vittinghoff (2004) posit that patients in different clusters can be assumed to be independent [34]. The shared frailty model with Gamma distributed frailties has been studied by Vaupel, et. al. (1979), Glidden and Vittinghoff (2004) [66, 34]. The advantage of using Gamma frailty model is that the marginal likelihood function usually has a closed form expression. However, the shared Gamma frailty model can only model positive association among multiple survival outcomes. McGilchrist and Aisbett (1991) used log-normal distribution for the frailty term [52]. By using multivariate log-normal frailties, it is possible to model negative association among multiple survival outcomes. A drawback of the log-normal frailty distribution discussed by Duchateau and Janssen (2008) is that the Laplace transform does not take a simple form [24]. Hougaard (1995) discussed the impact of other frailty distributions [38]. He used distributions like inverse Gaussian for the frailty term to overcome the restriction of the Gamma frailty that assumes the dependence is most important for late events.

Since random effects models involve high dimensional integrations and complex numerical computation, bootstrapping methods are frequently used for parameter estimation and inference in a frailty model. The approach of bootstrapping clustered data was first discussed in the design-based finite survey literature by Sitter (1992), Field and Welsh (2007) [62, 31]. Davison and Hinkley (1997) discussed the randomized cluster bootstrap [23]. Also, McCullagh (2000) considered the randomized cluster and
two-stage bootstraps and reverse two-stage bootstrap [51]. Andersson and Karlsson (2001) proposed a residual bootstrap method to reconstruct bootstrap observations from the bootstrap residuals [4].

In chapter 2, we had developed a bootstrap testing procedure to investigate the treatment-biomarker interaction for survival outcome. Due to the factor that multi-center studies are very common in modern clinical trials, it is of interest to extend these approaches to deal with the clustered survival outcome. In this chapter, the biomarker threshold model with an unknown biomarker cut-point and the random effect of clustering will be studied.

The objective of this chapter is to test the treatment-biomarker interaction for a biomarker threshold model with clustered survival outcome. In chapter 2, we had reviewed a lot of research in the literature about testing the biomarker-treatment interaction when dealing with independent survival outcomes. Under certain regularity conditions, the classic asymptotic results by Wilks(1938) and Serfling(1980) show that the likelihood ratio test statistic follows a chi-square limiting distribution [60, 71]. In the biomarker threshold model, the likelihood function is not a continuous function with respect to the threshold parameter. Therefore, the continuity and smooth conditions of the likelihood ratio test are no longer valid here. Since these regularity conditions are not satisfied for a biomarker threshold model with random effects and unknown cut-point for the biomarker variable, the ordinary likelihood ratio test cannot be directly applied.

In this chapter, we will develop a residual bootstrap resampling method to test with clustered survival outcome. The new test procedure is different from those by Loughin (1995) [50] and the one that we did in chapter 2 for independent survival
3.2. BIOMARKER THRESHOLD MODEL WITH RANDOM EFFECTS

outcomes. Here, we must take into account the special features of clustered data.

3.2 Biomarker threshold model with random effects

Statistical inferences for clustered survival outcome can be done through frailty models. Suppose that $n_i$ different patients are located in cluster $i$, for $i = 1, \ldots, m$. Let $z_{ij}, x_{ij}$ represent the treatment and the biomarker values for the $j^{th}$ patient in the $i^{th}$ cluster, respectively, for $i = 1, \ldots, m$, and $j = 1, \ldots, n_i$. Let $\tilde{T}_i = (\tilde{T}_{i1}, \ldots, \tilde{T}_{im})$ be the true failure times of the $n_i$ patients in cluster $i$, where $\tilde{T}_{ij}$ is the true failure time for the $j^{th}$ patient in the $i^{th}$ cluster. Let $C_{ij}$ denote the censoring time for the $j^{th}$ patient in the $i^{th}$ cluster. In general, we assume that $\tilde{T}_{ij}$ and $C_{ij}$ are independent given covariate $(x_{ij}, z_{ij})$. Also, let $\delta_{ij} = I(\tilde{T}_{ij} < C_{ij})$ represent the survival status indicator of the $j^{th}$ patient in the $i^{th}$ cluster, and $t_{ij} = \min(\tilde{T}_{ij}, C_{ij})$ is the observed failure/censoring time.

To associate the covariates and clustered survival time, we use a proportional hazards form for a biomarker threshold (Vaida and Xu (2000), Vaupel, et. al. (1979) [65, 66]),

$$\lambda_{ij}(t|w_i) = \lambda_0(t) \exp\{\beta_1 z_{ij} + \beta_2 I(x_{ij} > c) + \beta_3 z_{ij} I(x_{ij} > c) + w_i\}$$  \hspace{1cm} (3.1)

where $j = 1, \ldots, n_i$, $w = (w_1, \ldots, w_m)^T$ is the vector of random effects with $w_i$, the random effect for the $i^{th}$ cluster, for $i = 1, \ldots, m$. Suppose that $w \sim N_m(0, I\sigma^2)$, and $c$ is an unknown cut-point. We can view the $w_i$’s as a random sample from a density $f_\sigma(w)$, which is assumed to be a normal density. Here, $\beta_1$ is the main effect of the treatment, $\beta_2$ is the main effect of the biomarker, and $\beta_3$ is the treatment-biomarker
interaction effect. For \( i = 1, \ldots, m \), model (3.1) can be written as,

\[
\lambda_{ij}(t) = \lambda_0(t)v_i \exp\{\beta_1 z_{ij} + \beta_2 I(x_{ij} > c) + \beta_3 z_{ij} I(x_{ij} > c)\} \tag{3.2}
\]

where \( v_i = \exp(w_i) \) is called the frailty for the \( i^{th} \) cluster. Model (3.2) is the shared frailty model due to the fact that subjects in the same cluster all share the same frailty factor (Duchateau and Janssens (2008) [24]). A frailty corresponds to a random block effect that acts multiplicatively on the hazard rates of all subjects in a group. Note that for a biomarker with known cut-point \( c \), model (3.2) reduces to a typical Cox frailty model.

Here, we incorporate the random effect \( w_i \) in the model to describe heterogeneity among clusters. In some literature, \( w_i \) is called a random effect because it is used to model the between-cluster variation.

In general, one can integrate out the random effects to create the integrated partial likelihood function. For the Cox model with survival outcome, Ripatti and Palmgren (2000) showed that the integrated partial likelihood can be treated as a typical likelihood function, just as integrated full likelihoods [55]. For any fixed value of \( \beta = (\beta_1, \beta_2, \beta_3)^T, \sigma, \) and \( c \), the integrated partial likelihood for model (3.1) can be written as,

\[
IPL(\beta, \sigma, c) = \int_{-\infty}^{\infty} \prod_{i=1}^{m} PL_i(\beta, c | w)f_\sigma(w)dw, \tag{3.3}
\]

where \( f_\sigma(w) \) is the density function of a multivariate Normal distribution,

\[
f_\sigma(w) = \frac{1}{(2\pi)^{\frac{m}{2}}\sigma^{\frac{1}{2}}} \exp\left(-\frac{w^Tw}{2\sigma^2}\right),
\]
3.2. BIOMARKER THRESHOLD MODEL WITH RANDOM EFFECTS

and

\[ PL_i(\beta, c|w) = \prod_{j=1}^{n_i} \left[ \frac{\exp(\eta_{ij})}{\sum_{k=1}^{m} \sum_{h=1}^{n_k} Y_{kh}(t_{ij}) \exp(\eta_{kh})} \right]^{\delta_{ij}}, \]

in which \( \eta_{ij} = \beta_1 z_{ij} + \beta_2 I(x_{ij} > c) + \beta_3 z_{ij} I(x_{ij} > c) + w_i \) is the linear score for subject \( j \) in cluster \( i \). Here, \( Y_{ij}(t) \) describes the risk set, \( Y_{ij}(t) = 1 \) if subject \( j \) in the \( i^{th} \) cluster is still under observation at time \( t \) and 0 otherwise (Therneau and Grambsch (2000) [64]).

Let \( \ell = \ell(\beta, \sigma, c) \) be the logarithm of the integrated partial likelihood \( IPL \) in (3.3).

Note that, for a given \( c \), the maximum likelihood estimates of the parameters \( \beta, w = (w_i, \ldots, w_m)^T \), and \( \sigma \) cannot be obtained by directly maximizing \( IPL \) because the integral in (3.3) usually does not have analytical expression. Based on the results by Ripatti and Palmgren (2000), Therneau and Grambsch (2000), [55, 64], we can use the Laplace approximation for the integration in (3.3) in order to maximize \( \ell = \log(IPL) \) as follows,

\[ \ell = \ell(\beta, \sigma, c) = \log[IPL(\beta, \sigma, c)] \approx -\frac{1}{2} \log(\sigma^2) - \frac{1}{2} \log |G''(\tilde{w})| - G(\tilde{w}), \quad (3.4) \]

where

\[ G(w) = -\left[ \sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} \left\{ \eta_{ij} - \log \left( \sum_{k,h} Y_{kh}(t_{ij}) \exp(\eta_{kh}) \right) \right\} \right] - \frac{1}{2\sigma^2} \sum_{i=1}^{m} w_i^2, \]

where \( \tilde{w} \) is the solution to the partial derivatives of \( G(w) \) with respect to \( w \), i.e.

\[ G'(\tilde{w}) = \frac{\partial G(\tilde{w})}{\partial \tilde{w}} = 0, \]
and
\[ G''(w) = \frac{\partial^2 G(w)}{\partial w \partial w^T}|_{\hat{w}}. \]

If \( \sigma \) was known, and \( w \) was considered as fixed effects parameter, then \(-G\) would be a penalized partial likelihood,
\[
PPL(\beta, \sigma, w) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} \left\{ \eta_{ij} - \log \left( \sum_{k,h} Y_{kh}(t_{ij}) \exp(\eta_{kh}) \right) \right\} - \frac{1}{2\sigma^2} \sum_{i=1}^{m} w_i^2.
\]

Based on the results by Ripatti and Palmgren (2000), we can ignore the first two terms in (3.4) and maximize \( PPL \) to estimate \((\beta(\sigma), w(\sigma))\) for a fixed value of \( \sigma \). The procedure of calculating the maximum likelihood estimates of \( \beta \), \( w \), and \( \sigma \), for a given \( c \), can be summarized as follows. For a fixed \( \sigma \) and \( c \), the values of \( \beta \) and \( w \) that maximize \( PPL \) can be obtained easily by the same method as an ordinary Cox model. Therefore, choosing an initial value for \( \sigma \), we can write the first-order partial derivatives of \( PPL \) with respect to \( \beta \) and \( w \) as
\[
\frac{\partial}{\partial \beta_1} PPL(\beta, \sigma, w) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} \left[ z_{ij} - \frac{z_{ij} \exp(\eta_{ij})}{\sum_{k,h} Y_{kh}(t_{ij}) \exp(\eta_{kh})} \right],
\]
\[
\frac{\partial}{\partial \beta_2} PPL(\beta, \sigma, w) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} \left[ I(x_{ij} > c) - \frac{I(x_{ij} > c) \exp(\eta_{ij})}{\sum_{k,h} Y_{kh}(t_{ij}) \exp(\eta_{kh})} \right],
\]
\[
\frac{\partial}{\partial \beta_3} PPL(\beta, \sigma, w) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} \left[ z_{ij} I(x_{ij} > c) - \frac{z_{ij} I(x_{ij} > c) \exp(\eta_{ij})}{\sum_{k,h} Y_{kh}(t_{ij}) \exp(\eta_{kh})} \right],
\]
\[
\frac{\partial}{\partial w_i} PPL(\beta, \sigma, w) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} \left[ 1 - \frac{\exp(\eta_{ij})}{\sum_{k,h} Y_{kh}(t_{ij}) \exp(\eta_{kh})} \right] - \frac{1}{\sigma^2} w_i.
\]
Then, one can solve the equations by Newton-Raphson method to obtain \( [\hat{\beta}(\sigma), \hat{w}(\sigma)] \).
In the next step, we use these values in (3.4) to obtain \(\ell(\hat{\beta}(\sigma),\sigma,c)\), which is the approximated profile log-likelihood function for \(\sigma\), and maximize it to update the value of \(\sigma\). The procedure will be repeated until convergence. The procedure of estimating the parameters can be done by the package “coxme” in R software (Therneau (2015), [63]). Fitting a Cox frailty model using “coxme” package in R, we can get \(\hat{\beta}, \hat{w},\) and \(\hat{\sigma}\).

Our goal is to test the null hypothesis

\[ H_0 : \beta_3 = 0 \text{ vs } H_1 : \beta_3 \neq 0 \]

Define the null parameter space \(\Theta_0 = \{(\beta_1, \beta_2, \beta_3, \sigma, c) : (\beta_1, \beta_2) \in \mathbb{R}^2, \beta_3 = 0, \sigma \in \mathbb{R}^+, 0 < c < 1\}\), and the entire parameter space \(\Theta = \{(\beta_1, \beta_2, \beta_3, \sigma, c) : (\beta_1, \beta_2, \beta_3) \in \mathbb{R}^3, \sigma \in \mathbb{R}^+, 0 < c < 1\}\). For an unknown cut-point \(c\), the integrated partial likelihood function is not a continuous function with respect to \(c\). Therefore, the maximum likelihood estimate of \(c\) cannot be obtained directly. For any given \(c\), we define

\[ LR(c) = 2 \left\{ \ell(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\sigma}; c) - \ell(\tilde{\beta}_1, \tilde{\beta}_2, 0, \tilde{\sigma}; c) \right\}, \]

which follows a \(\chi^2_1\) distribution. Here, for a given value of \(c\), \(\tilde{\beta}\), and \(\hat{\beta}\) are the maximum likelihood estimates of \(\beta\), and \(\tilde{\sigma}\), \(\hat{\sigma}\) are the maximum likelihood estimates of \(\sigma\) under the null parameter space \(\Theta_{0c} = \{(\beta_1, \beta_2, \beta_3, \sigma) : (\beta_1, \beta_2) \in \mathbb{R}^2, \beta_3 = 0, \sigma \in \mathbb{R}^+\}\) and the entire parameter space \(\Theta_c = \{(\beta_1, \beta_2, \beta_3, \sigma) : (\beta_1, \beta_2, \beta_3) \in \mathbb{R}^3, \sigma \in \mathbb{R}^+\}\), for a given \(c\), respectively. However, since \(c\) is usually unknown, we define the test statistic for \(H_0\) as

\[ LR = \max_{0 < c < 1} LR(c). \quad (3.5) \]
Because the test statistic $LR$ is the maximum of $LR(c)$ over $0 < c < 1$, it does not follow the Chi-square distribution.

The exact distribution of the test statistic $LR$ in (3.5) is unknown in general. Similar to what we did in chapter 2, resampling methods can be used to approximate the asymptotic distribution of the proposed likelihood ratio test statistic $LR$. We will develop a method based on the residual bootstrap to identify the asymptotic distribution of $LR$, under the null hypothesis $H_0$ for clustered survival outcome.

### 3.3 The residual bootstrap method

Loughin (1995) proposed the residual bootstrap method to make statistical inference for regression parameters. In the residual bootstrap method, residuals follow a distribution that is independent of explanatory variables [50]. Therefore, unlike methods that bootstrap the individual subject, the explanatory variables do not need to be resampled, and they can be treated as some constants.

The residual bootstrap method is based on the fact that the partial likelihood function does not change under monotone increasing transformations of survival time $t$. Kalbfleisch and Prentice (1973) showed that, when the relative risk function is independent of time, and baseline hazard function $\lambda_0(t)$ is strictly positive over all open intervals, the partial likelihood function is invariant under monotone increasing transformations of time $t$ [41]. Furthermore, we can show that for the integrated partial likelihood, when the relative risk function is independent of time, and $\lambda_0(t)$ is strictly positive over all open intervals, the integrated partial likelihood function for the model (3.1) is invariant under monotone increasing transformations of time $t$.

Since the baseline cumulative distribution function is monotone in $t$, the integrated
partial likelihood remains the same either based on time \( t \) or based on the cumulative distribution function of time. Consequently, resampled residuals (estimated survival function) can be used to form observations from the cumulative distribution function based on which the test can be done. Furthermore, censoring can be incorporated in the resampling procedure.

The algorithm of residual bootstrap test (RBT) for the treatment-biomarker interaction with clustered survival outcome can be done in the following four steps.

- Estimate \( c \) by considering a grid of \( c \) values, \( 0 < c < 1 \). Obtain the maximum profile likelihood estimate of \( c \),

\[
\hat{c} = \arg \max_{0 < c < 1} \ell(\hat{\beta}_1 c, \hat{\beta}_2 c, \hat{\beta}_3 c, \hat{\sigma}_c; c),
\]

where \( \hat{\beta}_1 c, \hat{\beta}_2 c, \hat{\beta}_3 c, \) and \( \hat{\sigma}_c \) are the maximum likelihood estimates of \( \beta_1, \beta_2, \beta_3, \) and \( \sigma \) for each given value of \( c \). Then, taking \( c = \hat{c} \), fit the Cox frailty model (3.1) based on the original data \((t_{ij}, \delta_{ij}, x_{ij}), i = 1, \ldots, m, j = 1, \ldots, n_i,\) to obtain \( \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \) and \( \hat{\sigma} \). We can also obtain \( \hat{\mathbf{w}} = (\hat{w}_1, \ldots, \hat{w}_m)^T \) which is the estimated vector of random effects of the Cox frailty model fitted under the entire parameter space \( \Theta \). In the R package “coxme”, \( \hat{\mathbf{w}} \) is a part of the optimization procedure to estimate the maximum of the penalized partial likelihood function with Laplace transformation. The R syntax to obtain \( \hat{\mathbf{w}} \) is “\( \mathbf{w} = \text{random.effects} \)”. Here, \( \hat{c} \) is the maximum profile likelihood estimate of \( c \) on the parameter space \( \Theta \). Estimate survival function \( \hat{u}_{ij} = \hat{S}(t_{ij}) = \left[ \hat{S}_0(t_{ij}) \right]^{\hat{\eta}_{ij}}, \) for \( i = 1, \ldots, m, j = 1, \ldots, n_i, \) where \( \hat{\eta}_{ij} = \hat{\beta}_1 z_{ij} + \hat{\beta}_2 I(x_{ij} > \hat{c}) + \hat{\beta}_3 z_{ij} I(x_{ij} > \hat{c}) + \hat{\mathbf{w}}_i. \)
3.3. THE RESIDUAL BOOTSTRAP METHOD

- Generate the bootstrap data \((u_{ij}^*, \delta_{ij}^*, x_{ij})\) by sampling with replacement from \((\hat{u}_{ij}, \delta_{ij}, x_{ij})\), for \(i = 1, \ldots, m\) and \(j = 1, \ldots, n_i\). Note that we don’t resample the covariates here. Resampled time and censoring status pairs \((u_{ij}^*, \delta_{ij}^*)\) will be associated to the subject with \(x_{ij}\) in the original dataset, which is the value of covariate vector for the \(j^{th}\) patient in the \(i^{th}\) cluster. Calculate the probability-scale failure time, under the null hypothesis \(H_0\),

\[
y_{ij}^* = 1 - (u_{ij}^*) \exp\{-\{z_{ij} \tilde{\beta}_1 + I(x_{ij} > \tilde{c}) \tilde{\beta}_2\} + \tilde{w}_i\},
\]

where \(\tilde{\beta}_1\) and \(\tilde{\beta}_2\) are the maximum likelihood estimates, and \(\tilde{c}\) is the profile likelihood estimate of \(c\) on the null parameter space \(\Theta_0\). Also, \(\tilde{\mathbf{w}} = (\tilde{w}_1, \ldots, \tilde{w}_m)^T\) is the estimated vector of random effects for the fitted model under the null parameter space \(\Theta_0\), which can be obtained by fitting the model

\[
\lambda_{ij}(t) = \lambda_0(t) \exp\{\beta_1 z_{ij} + \beta_2 I(x_{ij} > c) + w_i\},
\]

and using “\(\tilde{\mathbf{w}} = \text{random.effects(fit0)}\)” in “coxme” package in R software to obtain the estimate of \(\mathbf{w}\) for each cluster.

- For each value of \(c\), fit two Cox frailty models under the parameter spaces \(\Theta_c\) and \(\Theta_{0c}\) for the bootstrapped data \((y_{ij}^*, \delta_{ij}^*, x_{ij})\), \(i = 1, \ldots, m, j = 1, \ldots, n_i\) (\(y^*\) as time and \(\delta^*\) as censoring indicator) to calculate the likelihood ratio statistic \(LR_b^*(c)\). Obtain the bootstrap version of the test statistic for the \(b^{th}\) bootstrap sample,

\[
LR_b^* = \max_{0 < c < 1} LR_b^*(c).
\]
• Repeat Steps 2, 3 for $B$ times. This gives $B$ replications of $LR^*_b$, for $b = 1, \ldots, B$.

This allows us to approximate the distribution of test statistic $LR$ given by (3.5). The p-value of the RBT is the proportion of $\{LR^*_b\}_{b=1}^B$ with values greater than the observed $LR$,

\[
P_{RBT} = \frac{1}{B} \sum_{b=1}^{B} I\{LR^*_b > LR\}, \tag{3.6}
\]

where $LR$ is the maximum likelihood ratio statistic of the form (3.5) calculated based on the original data.

3.4 Simulation studies

3.4.1 Simulation model design

To evaluate the finite sample performances of the residual bootstrap method for clustered survival outcome, we conducted a series of simulation studies. We further compared the residual bootstrap test to the permutation test (PT) proposed by Jiang et al. (2007) [39]. The idea of PT is to approximate the distribution of the test statistic by considering $B$ different permutations of assigning the treatments to the patients randomly. For clustered data, the permutation method can be applied by considering different permutations of assigning the treatments to the patients among different clusters. By doing this, we are implicitly assuming that there is no main treatment effects, which may be invalid in many situations.

For each parameter combination below, we repeated the simulation $R = 500$ times. Suppose that there are $m = 50$ clusters with $n_i = 6$ patients in each cluster, for $i = 1, \ldots, m$. To compare these two methods, we considered different combinations of $\beta_1, \beta_2, \beta_3 = \log(0.3), \log(0.5)$, and true cut-points $c_0 = 0.25, 0.6, \text{and } 0.75$ in the
survival outcomes were generated from a Weibull distribution with the conditional hazard function given by

\[ \lambda_{ij}(t|w_i) = \nu \gamma (\gamma t)^{\nu-1} \exp \{ \beta_1 z_{ij} + \beta_2 I(x_{ij} > c) + \beta_3 z_{ij} I(x_{ij} > c) + w_i \}, \]  

(3.7)

where \( w_i \)'s are independent identically distributed samples generated from \( N(0, (0.5)^2) \), for \( i = 1, \ldots, m \). We did simulations for \( \nu = 0.5, 1.0, 1.5, 2.0 \), and \( \gamma = 2 \). The biomarker values \( x_{ij} \) were generated from a Uniform distribution on the interval \((0,1)\). Censoring times were generated from a Uniform distribution \( U(0,3) \) such that the censoring rate is between 22%-32%.

Three different designs were used when generating the data set. These designs were different in how the treatments were assigned to the patients. The rational of using these three designs can be found in chapter 2. In design I, we assigned the new treatment to 80% of the patients, and the placebo control to 20% of the patients. In design II, if the biomarker value for a patient was greater than 0.5, then the treatment was assigned to the patient with 0.75 probability; otherwise, the treatment was assigned to the patient with 0.5 probability. In design III, we assigned two treatments to the patients with equal weights.

To calculate the test statistic for testing the null hypothesis, we used a grid of candidate biomarker cut-point values of 0.2,\ldots,0.8, increased by 0.02. We also can try a grid search on the interval \((0.05,0.95)\) increased by 0.01 by ignoring some errors in running the simulation program. When the cut-point \( c \) is near 0 or 1, the biomarker defined subset becomes too small such that some of the regression coefficients do not converge to a finite number. We address this issue by assigning a very small value
(e.g. $-n \times 10$) to the log likelihood corresponding such a threshold parameter $c$. We also run the simulations where the profile likelihoods for threshold parameter $c$ are calculated between 0.05 and 0.95 with step size equals to 0.01 and obtained similar results. Since the cut-point $c$ is unknown, we use $c = \hat{c}$ that maximizes the logarithm of the integrated partial likelihood function (Pons (2003)[54]). We will later show in simulation study that the profile likelihood method provides a consistent estimate of the threshold parameter $c$.

3.4.2 Simulation results

For the bootstrap and permutation methods, we used $B = 200$ bootstrap/permutation samples for each simulated data set. The results for the empirical test sizes are presented in Table 3.1. The empirical test size is the percentage of the simulated replications that reached the pre-specified level of the statistical significance when the null hypothesis is true. To calculate the empirical test size, clustered survival data must be generated under the null hypothesis $H_0$. If the empirical size of the test turns out to be close to the nominal significance level of $\alpha = 0.05$, the proposed test method can be considered to control type I error for testing the null hypothesis $H_0$. 
### 3.4. SIMULATION STUDIES

Table 3.1: The empirical size of the test for testing $H_0: \beta_3 = 0$ for clustered survival outcome$^1$.

| $\nu$ | $c_0$ | $\exp(\beta_1)$ | $\exp(\beta_2)$ | Design I
| PT | RBT | Design II
| PT | RBT | Design III
| PT | RBT |
|-----|-----|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 0.5 | 0.25 | 0.3 | 0.5 | 0.048 | 0.044 | 0.048 | 0.050 | 0.048 | 0.048 |
| 0.5 | 0.25 | 0.5 | 0.3 | 0.072 | 0.060 | 0.062 | 0.052 | 0.070 | 0.062 |
| 0.5 | 0.25 | 0.5 | 0.5 | 0.080 | 0.052 | 0.072 | 0.048 | 0.090 | 0.058 |
| 0.5 | 0.60 | 0.3 | 0.5 | 0.052 | 0.040 | 0.070 | 0.058 | 0.068 | 0.054 |
| 0.5 | 0.60 | 0.5 | 0.3 | 0.064 | 0.048 | 0.090 | 0.068 | 0.052 | 0.048 |
| 0.5 | 0.75 | 0.3 | 0.5 | 0.057 | 0.040 | 0.052 | 0.060 | 0.048 | 0.052 |
| 0.5 | 0.75 | 0.5 | 0.3 | 0.047 | 0.026 | 0.050 | 0.038 | 0.046 | 0.048 |
| 0.5 | 0.75 | 0.5 | 0.5 | 0.052 | 0.052 | 0.056 | 0.080 | 0.058 | 0.048 |
| 1.0 | 0.25 | 0.3 | 0.5 | 0.066 | 0.052 | 0.066 | 0.058 | 0.052 | 0.048 |
| 1.0 | 0.25 | 0.5 | 0.3 | 0.070 | 0.050 | 0.060 | 0.048 | 0.070 | 0.056 |
| 1.0 | 0.25 | 0.5 | 0.5 | 0.056 | 0.048 | 0.056 | 0.044 | 0.059 | 0.053 |
| 1.0 | 0.60 | 0.3 | 0.5 | 0.058 | 0.046 | 0.080 | 0.060 | 0.060 | 0.056 |
| 1.0 | 0.60 | 0.5 | 0.3 | 0.052 | 0.044 | 0.092 | 0.066 | 0.048 | 0.050 |
| 1.0 | 0.75 | 0.3 | 0.5 | 0.066 | 0.038 | 0.066 | 0.070 | 0.046 | 0.050 |
| 1.0 | 0.75 | 0.5 | 0.3 | 0.056 | 0.048 | 0.060 | 0.058 | 0.048 | 0.054 |
| 1.0 | 0.75 | 0.5 | 0.5 | 0.048 | 0.036 | 0.062 | 0.056 | 0.058 | 0.060 |
| 1.5 | 0.25 | 0.3 | 0.5 | 0.052 | 0.034 | 0.082 | 0.056 | 0.053 | 0.053 |
| 1.5 | 0.25 | 0.5 | 0.3 | 0.065 | 0.054 | 0.068 | 0.054 | 0.062 | 0.052 |
| 1.5 | 0.25 | 0.5 | 0.5 | 0.060 | 0.048 | 0.055 | 0.049 | 0.052 | 0.046 |
| 1.5 | 0.60 | 0.3 | 0.5 | 0.042 | 0.022 | 0.083 | 0.065 | 0.047 | 0.059 |
| 1.5 | 0.60 | 0.5 | 0.3 | 0.054 | 0.032 | 0.107 | 0.066 | 0.054 | 0.056 |
| 1.5 | 0.75 | 0.3 | 0.5 | 0.048 | 0.024 | 0.072 | 0.060 | 0.056 | 0.050 |
| 1.5 | 0.75 | 0.5 | 0.3 | 0.064 | 0.042 | 0.085 | 0.064 | 0.056 | 0.048 |
| 1.5 | 0.75 | 0.5 | 0.5 | 0.046 | 0.032 | 0.074 | 0.059 | 0.048 | 0.043 |
| 2.0 | 0.25 | 0.3 | 0.5 | 0.046 | 0.030 | 0.062 | 0.054 | 0.054 | 0.040 |
| 2.0 | 0.25 | 0.5 | 0.3 | 0.054 | 0.032 | 0.070 | 0.054 | 0.052 | 0.046 |
| 2.0 | 0.25 | 0.5 | 0.5 | 0.052 | 0.034 | 0.072 | 0.062 | 0.052 | 0.048 |
| 2.0 | 0.60 | 0.3 | 0.5 | 0.042 | 0.015 | 0.076 | 0.050 | 0.054 | 0.040 |
| 2.0 | 0.60 | 0.5 | 0.3 | 0.042 | 0.028 | 0.110 | 0.052 | 0.050 | 0.048 |
| 2.0 | 0.75 | 0.3 | 0.5 | 0.049 | 0.026 | 0.062 | 0.058 | 0.056 | 0.046 |
| 2.0 | 0.75 | 0.5 | 0.3 | 0.060 | 0.028 | 0.084 | 0.056 | 0.038 | 0.042 |
| 2.0 | 0.75 | 0.5 | 0.5 | 0.048 | 0.028 | 0.074 | 0.056 | 0.054 | 0.048 |

$^1$ Significance level $\alpha = 0.05$. Clustered failure time were generated based on the Weibull distribution in (3.7) with $\gamma = 2.0$, $m = 50$ clusters, $n_i = 6$ patients in each cluster, $B = 200$ bootstrap samples, $R = 500$ replications, under the null hypothesis.
As we see in Table 3.1, the RBT has the empirical test sizes around significance level $\alpha = 0.05$, while the PT does not control type I error in most cases, especially for a complicated design such as design II. For some combinations of the parameters, the empirical type I error based on the PT is two times larger than the nominal significance level, while the RBT works really well. For example, when $\nu = 1.5, c_0 = 0.6, \beta_1 = \log(0.5), \beta_2 = \log(0.3)$, the empirical test size for PT is 0.107, while the empirical test size for RBT is 0.066, using simulation design II. This is due to the fact that the PT automatically makes the model assumption that $\beta_1$ is 0, which may not be true in most scenarios. Furthermore, in many cases, especially using design I, the RBT is more conservative compared to the PT, with most of the empirical test sizes less than the nominal level of 0.05.

Moreover, we obtained the empirical power of the test for different values of $\beta_3 \neq 0$ for clustered survival outcome based on $R = 500$ replications. The empirical power of the test is the percentage of the simulated replications that reached the pre-specified level of the statistical significance when the alternative hypothesis is true. Simulation results for the empirical power of the test for RBT, using $m = 50$, were presented in Table 3.2. Since the PT method does not control type I error, it does not make sense to consider its test power.
3.4. SIMULATION STUDIES

Table 3.2: The empirical power of RBT under $H_1: \beta_3 \neq 0$, for clustered survival outcome with $m = 50^1$.

<table>
<thead>
<tr>
<th>$\nu$</th>
<th>$c_0$</th>
<th>$\exp(\beta_1)$</th>
<th>$\exp(\beta_2)$</th>
<th>$\exp(\beta_3)$</th>
<th>Design I</th>
<th>Design II</th>
<th>Design III</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.333</td>
<td>0.356</td>
<td>0.332</td>
</tr>
<tr>
<td>0.5</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
<td>0.157</td>
<td>0.322</td>
<td>0.298</td>
</tr>
<tr>
<td>0.5</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.301</td>
<td>0.386</td>
<td>0.415</td>
</tr>
<tr>
<td>0.5</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.245</td>
<td>0.364</td>
<td>0.485</td>
</tr>
<tr>
<td>0.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.438</td>
<td>0.598</td>
<td>0.684</td>
</tr>
<tr>
<td>0.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.229</td>
<td>0.415</td>
<td>0.510</td>
</tr>
<tr>
<td>1.0</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.357</td>
<td>0.471</td>
<td>0.457</td>
</tr>
<tr>
<td>1.0</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
<td>0.256</td>
<td>0.343</td>
<td>0.321</td>
</tr>
<tr>
<td>1.0</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.355</td>
<td>0.435</td>
<td>0.491</td>
</tr>
<tr>
<td>1.0</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.369</td>
<td>0.395</td>
<td>0.519</td>
</tr>
<tr>
<td>1.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.547</td>
<td>0.645</td>
<td>0.723</td>
</tr>
<tr>
<td>1.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.364</td>
<td>0.439</td>
<td>0.539</td>
</tr>
<tr>
<td>1.5</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.401</td>
<td>0.591</td>
<td>0.546</td>
</tr>
<tr>
<td>1.5</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
<td>0.276</td>
<td>0.388</td>
<td>0.365</td>
</tr>
<tr>
<td>1.5</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.404</td>
<td>0.514</td>
<td>0.619</td>
</tr>
<tr>
<td>1.5</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.416</td>
<td>0.501</td>
<td>0.645</td>
</tr>
<tr>
<td>1.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.659</td>
<td>0.798</td>
<td>0.872</td>
</tr>
<tr>
<td>1.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.435</td>
<td>0.586</td>
<td>0.709</td>
</tr>
<tr>
<td>2.0</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.407</td>
<td>0.625</td>
<td>0.587</td>
</tr>
<tr>
<td>2.0</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
<td>0.263</td>
<td>0.387</td>
<td>0.369</td>
</tr>
<tr>
<td>2.0</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.425</td>
<td>0.589</td>
<td>0.715</td>
</tr>
<tr>
<td>2.0</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.459</td>
<td>0.569</td>
<td>0.739</td>
</tr>
<tr>
<td>2.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.752</td>
<td>0.858</td>
<td>0.948</td>
</tr>
<tr>
<td>2.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.505</td>
<td>0.681</td>
<td>0.792</td>
</tr>
</tbody>
</table>

$^1$ Significance level $\alpha = 0.05$. Clustered failure time were generated based on the Weibull distribution in (3.7) with $\gamma = 2.0$, $m = 50$ clusters, $n_i = 6$ patients in each cluster, $B = 200$ bootstrap samples, $R = 500$ replications, under the alternative hypothesis.

For $m = 50$ clusters, the empirical powers do not reach 70% for many settings. We further conduct simulation studies for $m = 100$ clusters with all other parameters the same as Table 3.2. The results are reported in Table 3.3, and we observed that with 100 clusters, 6 patients per cluster, we have more than 70% power to reject the null hypothesis $H_0$ in most parameter settings. To have more powerful test, the simulations for $m = 150$ or 200 can be done. By considering the empirical power
values, we can conclude the RBT works well for testing the treatment-biomarker interaction effect when dealing with clustered survival outcome.

Table 3.3: The empirical power of RBT under $H_1 : \beta_3 \neq 0$, for clustered survival outcome with $m = 100^1$.

<table>
<thead>
<tr>
<th>$\nu$</th>
<th>$c_0$</th>
<th>$\exp(\beta_1)$</th>
<th>$\exp(\beta_2)$</th>
<th>$\exp(\beta_3)$</th>
<th>Design I</th>
<th>Design II</th>
<th>Design III</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.569</td>
<td>0.649</td>
<td>0.717</td>
</tr>
<tr>
<td>0.5</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
<td>0.367</td>
<td>0.452</td>
<td>0.529</td>
</tr>
<tr>
<td>0.5</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.528</td>
<td>0.564</td>
<td>0.698</td>
</tr>
<tr>
<td>0.5</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.521</td>
<td>0.493</td>
<td>0.704</td>
</tr>
<tr>
<td>0.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.701</td>
<td>0.757</td>
<td>0.810</td>
</tr>
<tr>
<td>0.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.479</td>
<td>0.618</td>
<td>0.646</td>
</tr>
<tr>
<td>1.0</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.664</td>
<td>0.775</td>
<td>0.847</td>
</tr>
<tr>
<td>1.0</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
<td>0.463</td>
<td>0.571</td>
<td>0.638</td>
</tr>
<tr>
<td>1.0</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.652</td>
<td>0.706</td>
<td>0.829</td>
</tr>
<tr>
<td>1.0</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.657</td>
<td>0.638</td>
<td>0.842</td>
</tr>
<tr>
<td>1.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.893</td>
<td>0.911</td>
<td>0.952</td>
</tr>
<tr>
<td>1.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.676</td>
<td>0.752</td>
<td>0.846</td>
</tr>
<tr>
<td>1.5</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.724</td>
<td>0.821</td>
<td>0.914</td>
</tr>
<tr>
<td>1.5</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
<td>0.491</td>
<td>0.616</td>
<td>0.712</td>
</tr>
<tr>
<td>1.5</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.742</td>
<td>0.801</td>
<td>0.927</td>
</tr>
<tr>
<td>1.5</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.777</td>
<td>0.744</td>
<td>0.925</td>
</tr>
<tr>
<td>1.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.958</td>
<td>0.977</td>
<td>1.000</td>
</tr>
<tr>
<td>1.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.779</td>
<td>0.867</td>
<td>0.954</td>
</tr>
<tr>
<td>2.0</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.721</td>
<td>0.859</td>
<td>0.929</td>
</tr>
<tr>
<td>2.0</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
<td>0.457</td>
<td>0.658</td>
<td>0.731</td>
</tr>
<tr>
<td>2.0</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.794</td>
<td>0.828</td>
<td>0.944</td>
</tr>
<tr>
<td>2.0</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.819</td>
<td>0.812</td>
<td>0.948</td>
</tr>
<tr>
<td>2.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.979</td>
<td>0.998</td>
<td>1.000</td>
</tr>
<tr>
<td>2.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.858</td>
<td>0.921</td>
<td>0.982</td>
</tr>
</tbody>
</table>

$^1$ Significance level $\alpha = 0.05$. Clustered failure time were generated based on the Weibull distribution in (3.7) with $\gamma = 2.0$, $m = 100$ clusters, $n_i = 6$ patients in each cluster, $B = 200$ bootstrap samples, $R = 500$ replications, under the alternative hypothesis.

To investigate the performance of the profile likelihood method in estimating unknown cut-point $c$, we calculated the empirical bias and standard error of the profile likelihood estimate under the alternative hypothesis. The results are presented
3.4. SIMULATION STUDIES

Table 3.4: Empirical bias and standard error of the profile likelihood estimate of $c$ for clustered survival outcome. The values were evaluated under $H_1 : \beta_3 \neq 0$, for design I. Failure time follows Weibull distribution in (3.7) with $\gamma = 2.0$, $m = 50$ clusters, $n_i = 6$ patients in each cluster, $B = 200$ bootstrap samples, $R = 500$ replications.

<table>
<thead>
<tr>
<th>$\nu$</th>
<th>$c_0$</th>
<th>$\exp(\beta_1)$</th>
<th>$\exp(\beta_2)$</th>
<th>$\exp(\beta_3)$</th>
<th>Bias($\hat{c}$)</th>
<th>S.E.($\hat{c}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.0014</td>
<td>0.024</td>
</tr>
<tr>
<td>1.0</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.0003</td>
<td>0.018</td>
</tr>
<tr>
<td>1.0</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>-0.0010</td>
<td>0.019</td>
</tr>
<tr>
<td>1.0</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>-0.0002</td>
<td>0.010</td>
</tr>
<tr>
<td>1.0</td>
<td>0.75</td>
<td>0.3</td>
<td>0.5</td>
<td>0.3</td>
<td>-0.0025</td>
<td>0.025</td>
</tr>
<tr>
<td>1.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>-0.0020</td>
<td>0.012</td>
</tr>
<tr>
<td>1.5</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.0002</td>
<td>0.017</td>
</tr>
<tr>
<td>1.5</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>-0.0023</td>
<td>0.015</td>
</tr>
<tr>
<td>1.5</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.0006</td>
<td>0.017</td>
</tr>
<tr>
<td>1.5</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.0001</td>
<td>0.008</td>
</tr>
<tr>
<td>1.5</td>
<td>0.75</td>
<td>0.3</td>
<td>0.5</td>
<td>0.3</td>
<td>0.0070</td>
<td>0.015</td>
</tr>
<tr>
<td>1.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>-0.0023</td>
<td>0.012</td>
</tr>
<tr>
<td>2.0</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.0010</td>
<td>0.018</td>
</tr>
<tr>
<td>2.0</td>
<td>0.25</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>-0.0020</td>
<td>0.013</td>
</tr>
<tr>
<td>2.0</td>
<td>0.60</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.0020</td>
<td>0.013</td>
</tr>
<tr>
<td>2.0</td>
<td>0.60</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.0004</td>
<td>0.008</td>
</tr>
<tr>
<td>2.0</td>
<td>0.75</td>
<td>0.3</td>
<td>0.5</td>
<td>0.3</td>
<td>-0.0020</td>
<td>0.014</td>
</tr>
<tr>
<td>2.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>-0.0030</td>
<td>0.011</td>
</tr>
</tbody>
</table>

Based on the result in Table 3.4, we observed that the empirical bias of the profile likelihood estimate for threshold parameter $c$ is very small. Therefore, we can conclude that the profile likelihood method provides a consistent estimate for the unknown cut-point $c$. 

1 The values were evaluated under $H_1 : \beta_3 \neq 0$, for design I. Failure time follows Weibull distribution in (3.7) with $\gamma = 2.0$, $m = 50$ clusters, $n_i = 6$ patients in each cluster, $B = 200$ bootstrap samples, $R = 500$ replications.
3.5. APPLICATION TO I-SPY 1 DATA SET

3.5 Application to I-SPY 1 data set

The I-SPY 1 clinical trial (Investigation of Serial Studies to Predict Your Therapeutic Response With Imaging and Molecular Analysis) was a multicenter breast cancer study designed to establish standards for collecting molecular and imaging data over the course of care (Esserman et al. (2012) [27]).

The I-SPY 1 trial was a collaboration of the American College of Radiology Imaging Network (ACRIN), Cancer and Leukemia Group B (CALGB), and the National Cancer Institute (NCI)s Specialized Programs of Research Excellence (SPORE). The data is available online from National Cancer Informatics Program (NCIP website): https://cbiit.nci.nih.gov/ncip/biomedical-informatics-resources/data.

This clinical trial enrolled 221 patients from ten different institutes (University of North Carolina at Chapel Hill, University of Chicago, University of California at San Francisco, Memory Sloan-Kettering Cancer Center, Georgetown University, University of Texas South Western Medical center, University of Pennsylvania, University of Washington, University of Alabama at Birmingham, and Eastern Cooperative Oncology Group (ECOG)). One of the study objectives is to investigate how pathologic Complete Response (pCR) affects the Disease Free Survival (DFS). The researchers are also interested in how pCR interacts with three biomarkers estrogen receptor total score (ER-TS), progesterone receptor status (PGR), and human epidermal growth factor receptor 2 (HER2). The histograms of these three biomarkers are presented in Figure 3.1. In this section, from three available biomarkers (ER-TS, PGR, and HER2), we choose estrogen receptor total score (ER-TS), as an example to illustrate how the proposed bootstrap method can be used for clustered survival outcome.

The ER total score in this data set is a measurement of all red score that ranges
3.5. APPLICATION TO I-SPY 1 DATA SET

from 0 to 8. Usually, an ER total score of 3 is used as cut-point in the application. However, here we assume that such a cut-point is unknown.

Here, we are interested in testing the interaction between pCR response and ER total score for this clustered survival data to see if patient’s DFS and pCR relationship depends on the ER total score or not, assuming unknown cut-point for the ER total score. Since ER total score (ER-TS) is an ordinary variable that takes integer value in 0 to 8, we will consider potential cut-point of $c = 0, 1, 2, 3, 4, 5, 6, 7$ for the indicator function $I(\text{ER-TS} > c)$.

Applying the proposed residual bootstrap method for clustered survival outcome data with $B = 2000$ bootstrap samples, we obtained a p-value=0.005 for the pCR

Figure 3.1: Histogram of ER-TS, PGR, and HER2 biomarkers for data from I-SPY 1 clinical trial.
and ER total score interaction, which is statistically significant. Therefore, we can say there is a significant interaction between pCR and ER total score, responding to the time to DFS. Figure 3.2 presents the profile integrated partial likelihood function vs. different values of $c$, under the alternative hypothesis. We identified the optimal cut-point of ER total score (based on the profile likelihood method) is equal to 3, which agrees with the value of cut-point used in clinical practice.

Figure 3.2: Profile integrated partial likelihood function vs. $c$ values, under $H_1$, for I-SPY 1 data set, with ER-TS as biomarker.

We further calculate hazard ratio (HR) of pCR versus no pCR for both subsets defined by ER-TS. For patients with ER-TS > 3, HR=1.754, 95% confidence interval=(0.482,6.38), p-value=0.394. For patients with ER-TS ≤ 3, HR=0.077, 95% confidence interval=(0.018,0.329), p-value=0.0005.
3.6 Discussions

In clinical trials and observational studies, it is possible that subjects may not be independent to each other. This generates clustered survival data. Moreover, we are interested to see if the new treatment benefits all patients in the same way or not. These lead us to develop biomarker threshold models with an unknown cut-point and random effect to test the treatment- biomarker interaction effect. However, the regularity conditions for the likelihood ratio test are not satisfied for this model. Therefore, the ordinary likelihood ratio test cannot be applied directly in this setting.

In this chapter, we proposed a resampling method to approximate the asymptotic

Figure 3.3: Kaplan-Meier estimates of disease-free survival for subsets defined by the biomarker ER-TS using \( \hat{c} = 3 \).

Moreover, Figure 3.3 presents the Kaplan-Meier estimates of disease-free survival for subsets defined by ER-TS using \( \hat{c} = 3 \). Considering both hazard ratio and Figure 3.3, we can say patients with pCR in ER-TS \( \leq 3 \) will have lower hazards.
distribution of the proposed likelihood ratio test statistic for the treatment-biomarker interaction effect using clustered survival outcome.

In numerical simulation studies, we showed that the proposed residual bootstrap method provides reasonable empirical test sizes in most settings. However, the permutation method cannot control type I error in some cases. To use the permutation method, the assumption $\beta_1 = 0$ must hold because the permutation of patients implicitly assumes no treatment main effects. This is a potential problem when using the permutation method, whereas the proposed residual bootstrap method has no such restrictions.

We applied the proposed residual bootstrap method for clustered data to a multicenter clinical trial I-SPY 1, and we showed that there is a significant interaction between pCR and ER total score on the DFS outcome. We also successfully identified a cut-point of ER total score=3 that happens to be the cut-point used in clinical practice.
Chapter 4

Penalized likelihood ratio test for binary outcome data

4.1 Overview

Suppose that we want to investigate the new treatment effect on the patients in different subsets defined by the biomarker. For example, one subset contains patients who may or may not benefit from the new treatment, and the other subset contains patients who benefit more from the new treatment. We are interested in testing the biomarker main effect and the treatment-biomarker interaction effect to describe different responses to the new treatment among these subsets. For a known cut-point, the test can be easily done by the fact that under the regularity and standard conditions, the likelihood ratio statistic has a Chi-square limiting distribution. But in many situations, a cut-point value to distinguish these two patient subsets has not been determined. The population has two mixture components (patient subsets) that respond differently to the new treatment. With the unknown biomarker cut-point, problems such as non-identifiability and irregularity arise as in the typical mixture model analysis. On the other hand, our focus of interests is on biomarker, treatment
effects and their interaction effect in a regression framework, which is different and more challenging than the typical mixture models.

Several authors like Bickel and Chernoff (1993), Chen (1995), Ghosh and Sen (1985), and Hartigan (1985) [7, 15, 33, 36] worked on problems when the regularity conditions for the likelihood ratio test are not satisfied, the test statistics may have complex limiting distributions or diverge to infinity. There are approaches to overcome the irregularities in mixture models or other models in the literature. Although these existing studies focused on typical mixture models with simple mixtures of densities, they did not consider mixture models induced by different regression models as in the treatment-biomarker inference problem we consider. Bohning, et. al. (1994) considered the distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family [8]. Cheng and Traylor (1995) studied the problem of non-regular maximum likelihood [17]. Davies (1977) studied hypothesis testing in the presence of a nuisance parameter [22]. Also, Feng and McCulloch (1996) applied bootstrap likelihood ratios to finite mixture models [29]. Lindsay (1989, 1995) considered mixture models [48, 49]. Chen, et. al. (1998) used the penalized likelihood ratio test for finite mixture models with multinomial observations [16]. Moreover, Chen, et. al. (2000) proposed a modified likelihood ratio test for homogeneity in the finite mixture models. They used the penalized likelihood ratio method for the test, and then derived the distribution of the proposed test statistic under the null and alternative hypotheses [14].

In this chapter, we will introduce a test statistic based on the penalized likelihood method to test both the biomarker main effect and the interaction effect between the treatment and the biomarker for unknown cut-point. The proposed test statistic
overcomes the non-identifiability and irregularity problems. Then, we will prove that the asymptotic distribution of the proposed penalized likelihood ratio statistic is a Chi-square distribution.

4.2 Model and notation

We consider a binary outcome variable $Y$, such that $Y = 1$ stands for death or disease, and $Y = 0$ for alive or no disease. Let $Z$ be the status of the treatments allocation for a patient with $Z = 0$ for control group and $Z = 1$ for treatment group. Let $X$ be the measurement of the biomarker, and without loss of generality, we assume $0 \leq X \leq 1$. Let $x$ denote the observed value of the biomarker. Suppose that patients with $x \leq c$ may not benefit from the new treatment or benefit less, whereas patients with $x > c$ benefit more from the new treatment, and $c$ is an unknown parameter. Suppose that we consider $n$ patients that are randomly assigned to the treatment and control groups. If patient $i$ belongs to the treatment group, then $z_i = 1$; otherwise, $z_i = 0$. Let $0 \leq x_i \leq 1$ represents the biomarker value for the $i^{th}$ patient. We use the following logistic regression model to investigate the association between the response variable $Y$, the treatment $Z$, and the biomarker variable $X$,

\[
\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 z_i + \beta_2 I(x_i > c) + \beta_3 z_i I(x_i > c), \ i = 1, \ldots, n, \tag{4.1}
\]

where $\beta_1$ is the main treatment effect, $\beta_2$ is the main biomarker effect, and $\beta_3$ is the treatment by biomarker interaction effect. The analysis is based on the model assumption that $Y_i \sim \text{Bernoulli}(1, p_i)$, where $p_i = P(Y_i = 1; x_i, z_i)$ is the probability of death.

The cut-point $c$ defines two components of the population which are two subsets
of the patients. The first subset contains the patients who may or may not benefit from the new treatment, and the other one contains the patients who benefit more from the new treatment. Therefore, model (4.1) can be viewed as a mixture model when the cut-point \( c \) is unknown, but defined through a mixture of two regression models.

We are interested in testing the null hypothesis, \( H_0 : \beta_2 = \beta_3 = 0 \) vs. the alternative hypothesis \( H_1 \): at least one of \( \beta_2 \) or \( \beta_3 \) is not equal to 0, i.e., whether there is a significant biomarker main effect or a treatment-biomarker interaction effect or not.

When \( c \) is known, the regularity conditions hold, and the likelihood ratio test statistic follows a Chi-square distribution with 2 degrees of freedom. Therefore, we can conduct the test based on the likelihood ratio method.

When the biomarker cut-point \( c \) is unknown, the null hypothesis can be written as \( H_0 : c = 0 \) or \( 1 \) vs. the alternative hypothesis \( H_1 : c \neq 0, 1 \). This is because either \( \beta_2 \neq 0 \) or \( \beta_3 \neq 0 \) implies that there exists a cut-point \( c, 0 < c < 1 \), that divides the patient population into two subsets with \( x \leq c \) or \( x > c \), giving two different regression model forms in (4.1). However, the regularity conditions are not satisfied for this model under \( H_0 \) because \( c = 0 \) or \( 1 \) are on the boundaries of the parameter space regarding to \( c \). In addition, model (4.1) is non-identifiable under \( H_0 \), as any value of \( c \) gives the same model form when \( \beta_2 = \beta_3 = 0 \).

Another challenge is that we cannot find the maximum likelihood estimate of \( c \) directly since \( I(x > c) \) is not continuous in \( c \). Therefore, we look for a continuous function to approximate \( I(x > c) \) and conduct the hypothesis testing. We found that the indicator function \( I(x > c) \) can be approximated by the continuous function
4.2. MODEL AND NOTATION

\[ w(x, c) = \frac{\exp\{K \cdot (x - c)\}}{1 + \exp\{K \cdot (x - c)\}}, \]

where \( K \) is a constant. When \( K \) is large (e.g., \( K = 100 \)), we have \( w(x, c) \approx I(x > c) \). This helps us to use a smooth function instead of the indicator function to estimate the unknown parameter \( \theta = (\beta_0, \beta_1, \beta_2, \beta_3, c)^T \), and we can simply use Newton-Raphson method.

Figure 4.1 compares function \( w(x, c) \) to \( I(x > c) \) for different \( K = 10, 50, 100, 200, 300, 500 \), and \( c = 0.5 \). Note that for larger \( K \), say \( K \geq 100 \), \( w(x, c) \) can provide a good approximation for \( I(x > c) \).

Figure 4.1: \( w(x, c), I(x > c) \) vs. \( x \), for \( c = 0.5, K = 10, 50, 100, 200, 300, 500 \).

Note that, by using \( w(x, c) \) instead of \( I(x > c) \), model (4.1) can be written as

\[
\log\left(\frac{p}{1 - p}\right) = \beta_0 + \beta_1 z + \beta_2 w(x, c) + \beta_3 z w(x, c),
\]

(4.2)
4.3. PENALIZED LIKELIHOOD RATIO TEST

which is an approximate expression for the mixture of regression models described by (4.1). Model (4.2) still has model irregularities that does not allow us to use ordinary likelihood ratio test statistic for testing $H_0$. The first problem remains that $c$ is located on the parameter space boundaries. In the model (4.2), the parameter $c$ takes its value on $[0,1]$. The situation with $c = 0$ or 1 implies that the patient population does not divide into two subsets defined by the biomarker, i.e., all patients benefit in the same way from the new treatment (which corresponds to the situation under the null hypothesis, when there is no main biomarker effect nor interaction effect between the biomarker and the treatment). In other words, when $\beta_2 = \beta_3 = 0$, then $c = 0$ or 1. The second problem for (4.1), the non-identifiability, still persists for model (4.2). We have to apply a method to overcome these irregularities in order to test for $H_0$.

4.3 Penalized likelihood ratio test

In this section, we introduce a penalized likelihood ratio test statistic to test the null hypothesis in the logistic regression model (4.2). In Section 4.3.1, we state some regularity conditions for theoretical development of the penalized likelihood ratio test. The theoretical properties of the maximum penalized likelihood method will be studied in Section 4.3.2. Also, by proving some lemmas and theorems, we prove that the limiting distribution of the penalized log-likelihood ratio statistic for testing $H_0 : \beta_2 = \beta_3 = 0$ is a Chi-square distribution with 3 degrees of freedom under $H_0$, with the extra one degree of freedom coming from the unknown parameter $c$. Testing for $H_0 : \beta_2 = \beta_3 = 0$ in model (4.1) or similarly in model (4.2), we have the following
4.3. PENALIZED LIKELIHOOD RATIO TEST

model under the null hypothesis,

$$H_0 : \log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 z.$$ 

We define the penalized log-likelihood function $p\ell_n$ as,

$$p\ell_n(\theta) = \ell_n(\theta) + g_n(c, u) = \sum_{i=1}^{n} \left[ \log f(y_i|\theta) + \log \left\{ 2c^{u_i}(1 - c)^{(1-u_i)} \right\} \right]$$

$$= \sum_{i=1}^{n} \left\{ \log f(y_i|\theta) + g(c, u_i) \right\},$$

(4.3)

where $g(c, u_i) = \log \left\{ 2c^{u_i}(1 - c)^{(1-u_i)} \right\}$, $g_n(c, u) = \sum_{i=1}^{n} g(c, u_i)$. Also,

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f(Y_i|\theta),$$

is the log-likelihood function for model (4.2), where $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, c)^T$. Here,

$$f(Y_i|\theta) = p_i^{y_i}(1 - p_i)^{(1-y_i)}$$

is based on the Bernoulli distribution for $Y_i$ with

$$p_i = \frac{\exp \left\{ \beta_0 + \beta_1 z_i + \beta_2 w(x_i, c) + \beta_3 z_i w(x_i, c) \right\}}{1 + \exp \left\{ \beta_0 + \beta_1 z_i + \beta_2 w(x_i, c) + \beta_3 z_i w(x_i, c) \right\}}.$$

Note that $p = p(\theta)$, where $\theta = (\beta, c)$ is a function of unknown parameters $\beta_0, \beta_1, \beta_2, \beta_3, c$. Hereafter, we will denote it by $p$.

We propose using $g_n(c, u) = \sum_{i=1}^{n} \log \left\{ 2c^{u_i}(1 - c)^{(1-u_i)} \right\}$ in (4.3) as a penalty term, in which $u = (u_1, .., u_n)^T$ is a random sample from Bernoulli distribution with parameters $(1,0.5)$. In fact, we penalize and prevent the model fit with $c$ close to 0 or 1 by adding the penalty term $g_n(c, u)$ to the log-likelihood function. This is a way to get over the non-identifiability problem of model (4.2) and guarantees that the maximum
penalized likelihood estimates of $\beta_2$ and $\beta_3$ converge to their true values under $H_0$ (will be shown in Section 4.3.2). The penalty term $g_n(c, u)$ has random components, which is different from the fixed penalty function in typical mixture model analysis. The new idea of random penalty term turns out to be very useful and convenient for proving the following Theorem 3 and Theorem 4.

When $\beta_2 = \beta_3 = 0$, model (4.2) (or (4.1)) has the same form when the cut-point parameter $c$ takes any value in $(0, 1)$. We write the null hypothesis $H_0$ as: $\beta_2 = \beta_3 = 0; c = \frac{1}{2}$. In general, the penalty term in (4.3) takes a non-positive value, which is maximized at $c^* = \bar{u}$. It is obvious that $\bar{u}$ converges in probability to 0.5. Therefore, we specify $c = \frac{1}{2}$ in $H_0$, for which the penalty term equals to 0.

Let $\Theta = \mathbb{R}^4 \times (0, 1)$ be the entire parameter space of $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, c)^T$, and $\Theta_0 = \mathbb{R}^2 \times \{0\} \times \{0\} \times \{0.5\}$ be the parameter space of $\theta$ under the null hypothesis $H_0$, with $\beta_2 = \beta_3 = 0$ and $c = \frac{1}{2}$.

We can apply the profile likelihood method to find the maximum profile penalized likelihood estimates of $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, c)^T$ on the entire parameter space $\Theta$. Based on the profile likelihood method,

$$\hat{c}_P = \arg \max_{0 < c < 1} p\ell_n(\hat{\beta}_c, c)$$

where $\hat{\beta}_c$ is the maximum penalized likelihood estimate of $\beta$ for a given value of $c$, and $p\ell_n$ is the log penalized likelihood function in (4.3). Then, we can use $\hat{c}_P$ to find the maximum penalized likelihood estimate of $\beta$ by maximizing $p\ell_n(\beta, \hat{c}_P)$ with respect to $\beta$.

The profile likelihood estimation can be used for both model (4.1) with $I(x > c)$, and model (4.2) with $w(x, c)$. However, we can also apply the Newton-Raphson
method to estimate the unknown parameter $\theta$ in model (4.2) and conduct a test for testing the null hypothesis $H_0$ based on the maximum penalized likelihood estimates of $\theta$. Newton-Raphson method can be used because we adopt model (4.2) which replaces $I(x > c)$ by a smooth function $w(x, c)$. It can be shown that both profile likelihood and Newton-Raphson methods return the same estimates for the parameter $\theta$.

The corresponding penalized log-likelihood ratio statistic for testing $H_0$ can be defined as

$$R_n = 2\{p\ell_n(\hat{\theta}) - p\ell_n(\tilde{\theta})\},$$

(4.4)

where $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{c})^T$ is the maximum penalized likelihood estimate on the entire parameter space $\Theta$ based on the Newton-Raphson method, and $\tilde{\theta} = (\tilde{\beta}_0, \tilde{\beta}_1, 0, 0, 0.5)^T$ is the maximum penalized likelihood estimate (which is the same as the maximum likelihood estimate) on the parameter space $\Theta_0$. Because $R_n$ involves the term from penalized likelihood function, classical asymptotic results for likelihood ratio test cannot be applied here directly. In the next step, we will derive the limiting distribution of $R_n$, under $H_0$.

Before studying the limiting distribution of the penalized likelihood ratio test statistic defined in (4.4), we will give some conditions and lemmas.

### 4.3.1 Regularity conditions

We need some regularity conditions to derive the asymptotic distribution of the penalized likelihood ratio test statistic.

C0) Parameters $\beta$ and $c$ must be interior points such that $0 < c < 1$, but $\beta$ can be any value.
4.3. PENALIZED LIKELIHOOD RATIO TEST

C1) Smoothness: The function \( f = f(y_i|\theta) \) has common support and is twice continuously differentiable with respect to \( \beta_0, \beta_1, \beta_2, \beta_3, \) and \( c. \) The first two derivatives are denoted by partial derivatives \( \frac{\partial f}{\partial \beta_j}, \frac{\partial^2 f}{\partial \beta_j^2}; \) \( j = 0, 1, 2, 3, \) and \( \frac{\partial f}{\partial c}, \frac{\partial^2 f}{\partial c^2}. \)

C2) The model under consideration is finitely identifiable if for any \( 0 < c_1, c_2 < 1, \)

\[
f(y_i|\beta, c_1) = f(y_i|\beta, c_2),
\]

for all \( y_i, \) implies \( c_1 = c_2, \) for all \( (\beta_0, \beta_1, \beta_2, \beta_3, i = 1, \ldots, n. \)

C3) Uniform strong law condition of large numbers. By this condition,

\[
n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(Y_i|\theta) \xrightarrow{a.s.} E\left\{ \frac{\partial}{\partial \theta} \log f(Y_1|\theta) \right\}, \quad \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial c} g(c, u_i) \xrightarrow{a.s.} E\left\{ \frac{\partial}{\partial c} g(c, u_1) \right\},
\]

where \( g(c, u_i) = \log\{2c^u(1 - c)^{1-u_i}\}, \)

\[
n^{-1} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta \partial \theta^T} \log f(Y_i|\theta) \xrightarrow{a.s.} E\left\{ \frac{\partial^2}{\partial \theta \partial \theta^T} \log f(Y_1|\theta) \right\}, \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial c^2} g(c, u_i) \xrightarrow{a.s.} E\left\{ \frac{\partial^2}{\partial c^2} g(c, u_1) \right\},
\]

uniformly in \( \theta \) (Ferguson (1996) and Rubin (1956) [30, 58]).

C4) This condition states the regularity condition needed for asymptotic properties of the maximum penalized likelihood estimates of the parameter \( \theta \) as

\[
\frac{\partial}{\partial \theta} \iint_{A} \{f(y|\theta) + g(c, u)\} dy du = \iint_{A} \frac{\partial}{\partial \theta} \{f(y|\theta) + g(c, u)\} dy du,
\]

\[
\frac{\partial^2}{\partial \theta_i \partial \theta_j} \iint_{A} \{f(y|\theta) + g(c, u)\} dy du = \iint_{A} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \{f(y|\theta) + g(c, u)\} dy du,
\]
for \(i, j = 1, \ldots, 5\), and \(A\) is the support of \(f(y|\theta) + g(c, u)\), i.e. \(A = \{(y, u) : f(y|\theta) > 0, g(c, u) > 0\}\). The \(\int\) would be replaced by \(\sum\) if \(Y\) is discrete.

C5) Let \(I(\theta)\) be the Fisher information matrix of \(p_{\ell n}(\theta), \theta \in \Theta\). \(I(\theta)\) is positive-definite for \(\theta\) in a small neighbourhood around \(\theta^{(0)}\), where \(\theta^{(0)}\) is the true value of \(\theta\).

It can be easily checked that for the term in penalized likelihood function,

\[
\log f(Y_i|\theta) + g(c, u_i) = \log\{p^{y_i}(1 - p)^{(1-y_i)}\} + \log\{2c^{u_i}(1 - c)^{(1-u_i)}\},
\]

the conditions C0-C5 hold.

### 4.3.2 Properties of the maximum penalized likelihood estimates

Suppose that \(\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{c})^T\) maximizes the penalized log-likelihood function defined in (4.3) on the entire parameter space \(\Theta\). We would like to prove that, under \(H_0\), \(\hat{c}\) is away from 0 or 1. Also, by proving Theorem 2, we will show that \(\hat{\theta}\) is consistent and converges to its true value when \(H_0\) is true.

**Lemma 1.** Under conditions C0- C5, and under the null hypothesis \(H_0\),

\[g_n(\hat{c}, u) = O_p(1)\]

**Proof.** Let \(R_n\) be the penalized likelihood ratio statistic defined in (4.4), and

\[M_n = 2\{\ell_n(\tilde{\theta}) - \ell_n(\tilde{\theta})\}\]

be the ordinary likelihood ratio statistic, where \(\tilde{\theta}\) is the maximum likelihood estimate
4.3. PENALIZED LIKELIHOOD RATIO TEST

of $\theta$ on $\Theta$. Since $g_n(c, u) \leq 0$ and $R_n \leq 2\{\ell_n(\hat{\theta}) - \ell_n(\tilde{\theta})\} \leq M_n$, then $0 \leq R_n \leq M_n$. Notice that on $\Theta_0$, $p\ell_n(\theta) = \ell_n(\theta)$, and the maximum likelihood estimate is $\tilde{\theta}$. Also, $M_n = O_p(1)$ (Chen and Chen (2001) [12]), so $R_n = O_p(1)$. Since $M_n$ is the maximum likelihood ratio statistic,

$$0 \leq R_n - g_n(\hat{c}, u) \leq M_n.$$ 

Therefore, $R_n - g_n(\hat{c}, u) = O_p(1)$, and $g_n(\hat{c}, u) = O_p(1)$. 

Based on Lemma 1, $\hat{c}$ must be bounded away from 0 or 1. This means there is a constant $\delta$, $\delta > 0$, such that $\hat{c}$ takes values in the interval $(\delta, 1 - \delta)$. Lemma 1 indicates that the penalized likelihood method guarantees that the estimate of $c$ does not occur on the boundaries of the parameter $c$ ($c = 0$ or 1). Recall that in the regularity conditions, $c$ must be an interior point such that $0 < c < 1$.

**Theorem 2.** Assume that Conditions C0-C5 hold. Let $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{c})^T$ be the maximum penalized likelihood estimate of $\theta$ on $\Theta$. Under $H_0$, $\hat{\theta}$ converges in probability to $\theta(0) = (\beta_0^{(0)}, \beta_1^{(0)}, 0, 0, 0.5)^T$ as $n \to \infty$.

**Proof.** We prove that $\hat{\theta}$ converges to $\theta(0)$ by contradiction. First, suppose one element of $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$, say $\hat{\beta}_3$, does not converge to its true value under $H_0$. There are positive values $\delta'$, $\epsilon'$ such that for a subsequence of $n$,

$$P(|\hat{\beta}_3| > \delta') > \epsilon'.$$

On the other hand, we can find $\delta > 0$ by Lemma 1 such that

$$P(\hat{c} \leq \delta \text{ or } \hat{c} \geq 1 - \delta) < \frac{\epsilon'}{2}.$$
Define \( A_n = \{ |\hat{\beta}_3| > \delta', \delta < \hat{c} < 1 - \delta \} \). Therefore, \( P(A_n) \geq \frac{\epsilon'}{2} \), for a subsequence of \( n \). When \( A_n \) occurs,

\[
\frac{p\ell_n(\hat{\theta})}{n} - \frac{p\ell_n(\theta^{(0)})}{n} \leq \frac{\ell_n(\hat{\theta}) - \ell_n(\theta^{(0)})}{n} \leq \sup_{|\beta_3| > \delta'} \frac{\ell_n(\beta, \hat{c}) - \ell_n(\theta^{(0)})}{n}, \tag{4.5}
\]

where \( \beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T \). Notice that

\[
\frac{\ell_n(\beta, \hat{c}) - \ell_n(\theta^{(0)})}{n} < 0, \tag{4.6}
\]

when \( \beta \) is bounded away from its true value \( \beta^{(0)} = (\beta_0^{(0)}, \beta_1^{(0)}, 0, 0)^T \). To see this, let \( \tilde{\beta}_c \) be the maximum likelihood estimate of \( \beta \) for a given \( c \) value, \( \delta < \hat{c} < 1 - \delta \), that maximizes \( \ell_n(\beta, c) \). Write

\[
\frac{\ell_n(\beta, \hat{c}) - \ell_n(\theta^{(0)})}{n} = \frac{\ell_n(\tilde{\beta}_c, \hat{c}) - \ell_n(\theta^{(0)}) - [\ell_n(\tilde{\beta}_c, \hat{c}) - \ell_n(\beta, \hat{c})]}{n}. \tag{4.7}
\]

Under \( H_0 \), the likelihood function has the same form for any given \( c, \delta < \hat{c} < 1 - \delta \). The estimate \( \tilde{\beta}_c \) has the maximum likelihood estimate properties. That is, for a given \( c \) value, \( \tilde{\beta}_c \overset{p}{\to} \beta^{(0)} \), for \( \delta < \hat{c} < 1 - \delta \). Moreover,

\[
\frac{\ell_n(\tilde{\beta}_c, \hat{c}) - \ell_n(\theta^{(0)})}{n} = \frac{(\tilde{\beta}_c - \beta^{(0)})^T J(\tilde{\beta}_c)(\tilde{\beta}_c - \beta^{(0)})}{n} + o_p(1) \tag{4.8}
\]

\[
= (\tilde{\beta}_c - \beta^{(0)})^T J_1(\beta^{(0)})(\tilde{\beta}_c - \beta^{(0)}) + o_p(1),
\]

where \( J(\tilde{\beta}_c) \) is the observed information matrix of \( \ell_n(\beta, \hat{c}) \) at the maximum likelihood estimate \( \tilde{\beta}_c \), and \( J_1(\beta^{(0)}) = -E \left[ \frac{\partial^2}{\partial \beta \partial \beta^T} \log f(Y_1|\beta, \hat{c}) \right] \).
4.3. PENALIZED LIKELIHOOD RATIO TEST

It follows from $\tilde{\beta}_c \overset{p}{\to} \beta^{(0)}$, that

$$\frac{\ell_n(\tilde{\beta}_c, \hat{c}) - \ell_n(\theta^{(0)})}{n} \overset{p}{\to} 0. \quad (4.9)$$

With similar derivation,

$$\frac{\ell_n(\tilde{\beta}_c, \hat{c}) - \ell_n(\beta, \hat{c})}{n} = (\tilde{\beta}_c - \beta)^T J_1(\beta^{(0)})(\tilde{\beta}_c - \beta) + o_p(1). \quad (4.10)$$

In $A_n$, $|\beta_3| > \delta'$, i.e. $\beta$ is bounded away from the true value $\beta^{(0)}$. We can find $\delta_1 > 0$ such that $||\tilde{\beta}_c - \beta|| > \delta_1$, where $||\tilde{\beta}_c - \beta|| = \max\{\tilde{\beta}_{c,j} - \beta_j; j = 0, 1, 2, 3\}$. By (4.10), we then have

$$\frac{\ell_n(\tilde{\beta}_c, \hat{c}) - \ell_n(\beta, \hat{c})}{n} > 0. \quad (4.11)$$

Combining (4.9), (4.11) with (4.7) gives (4.6).

This implies

$$P\left\{\frac{p\ell_n(\hat{\theta})}{n} - \frac{p\ell_n(\theta^{(0)})}{n} < 0\right\} \geq P(A_n) \geq \frac{\epsilon'}{2},$$

which contradicts the fact that $\hat{\theta}$ maximizes the penalized log-likelihood function. Hence, $\hat{\beta}_3 \overset{p}{\to} \beta^{(0)}_3 = 0$. The same argument applies to $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$. Therefore, we have $\hat{\beta} \overset{p}{\to} \beta^{(0)}$.

Next, we show that $\hat{c} \overset{p}{\to} 0.5$. Notice that

$$\frac{1}{n} g_n(\hat{c}, u) \overset{p}{\to} 0, \quad (4.12)$$
by Lemma 1, where

\[
\frac{1}{n} g_n(c, u) = \frac{1}{n} \sum_{i=1}^{n} \log \{2c^{u_i} (1 - c)^{(1 - u_i)} \} = \log 2 + \bar{u} \log c + (1 - \bar{u}) \log (1 - c).
\]

Suppose \( \frac{1}{n} g_n(c, u) \) is maximized at \( c = c^* \). It is easy to find that \( c^* = \bar{u} \), which satisfies

\[
\frac{1}{n} \frac{dg_n(c^*, u)}{dc} = 0.
\]

By Taylor expansion,

\[
\frac{1}{n} g_n(\hat{c}, u) = \frac{1}{n} g_n(c^*, u) + \frac{1}{n} \frac{dg_n(c^*, u)}{dc} (\hat{c} - c^*) + \frac{1}{n} \frac{d^2 g_n(c^*, u)}{dc^2} (\hat{c} - c^*)^2 + o_p(1)
\]

\[
= \log 2 + \bar{u} \log \bar{u} + (1 - \bar{u}) \log (1 - \bar{u}) + \left\{ -\frac{\bar{u}}{\bar{u}^2} - \frac{1 - \bar{u}}{(1 - \bar{u})^2} \right\} (\hat{c} - \bar{u})^2 + o_p(1).
\]

Note that \( \frac{dg_n(c^*, u)}{dc} = 0 \). Since \( \bar{u} \mathop\to\limits^P 0.5 \), we see that

\[
\frac{1}{n} g_n(\hat{c}, u) = -4(\hat{c} - \bar{u})^2 + o_p(1).
\]

This along with (4.12) implies that

\[
\hat{c} - \bar{u} \mathop\to\limits^P 0, \text{ i. e. } \hat{c} \mathop\to\limits^P 0.5.
\]
4.3. PENALIZED LIKELIHOOD RATIO TEST

4.3.3 Asymptotic distribution of the maximum penalized likelihood estimate \( \hat{\theta} \)

**Theorem 3.** Let \( \hat{\theta} \) be the maximum penalized likelihood estimate of \( \theta \) on the entire parameter space \( \Theta \), based on the penalized log-likelihood in (4.3) for model (4.2). Assume that conditions C0-C5 hold. Then,

1. \[
\sqrt{n} \left( \hat{\theta} - \left[ \theta + I_1^{-1}(\theta)E_\theta \{ p\ell'(Y_1, \theta) \} \right] \right) \xrightarrow{D} N_5 \left( 0, I_1^{-1}(\theta)Var_\theta \{ p\ell'(Y_1, \theta) \} I_1^{-1}(\theta) \right),
\]
   where \( p\ell(Y_1, \theta) = \log \left\{ f(y_1|\theta) \right\} + g(c, u_1) \) for \( g(c, u_1) \) defined above, and \( I_1(\theta) = E \left[ \{ p\ell'(Y_1, \theta) \} \{ p\ell'(Y_1, \theta) \}^T \right] \), for \[
p\ell'(Y_1, \theta) = \left( \frac{\partial p\ell(Y_1, \theta)}{\partial \beta_0}, \frac{\partial p\ell(Y_1, \theta)}{\partial \beta_1}, \frac{\partial p\ell(Y_1, \theta)}{\partial \beta_2}, \frac{\partial p\ell(Y_1, \theta)}{\partial \beta_3}, \frac{\partial p\ell(Y_1, \theta)}{\partial c} \right)^T.
\]
2. When \( H_0 \) is true, and \( \theta^{(0)} \) is the true value of \( \theta \),
\[
\sqrt{n} \left( \hat{\theta} - \theta^{(0)} \right) \xrightarrow{D} N_5 \left( 0, I_1^{-1}(\theta^{(0)}) \right).
\]

**Proof.** By (4.3), the score function (a vector in \( \mathbb{R}^5 \)) of the penalized log-likelihood is
\[
p\ell_n(\theta) = \sum_{i=1}^n p\ell(y_i, \theta) = \sum_{i=1}^n \left\{ \frac{\partial}{\partial \theta} \log f(y_i|\theta) + \frac{\partial}{\partial \theta} g(c, u_i) \right\} \quad (4.13)
\]
By Central Limit Theorem,
\[
\sqrt{n} \left[ \frac{1}{n} \sum_{i=1}^n p\ell'(Y_i, \theta) - E_\theta \{ p\ell'(Y_1, \theta) \} \right] \xrightarrow{D} N_5 \left( 0, Var_\theta \{ p\ell'(Y_1, \theta) \} \right), \quad (4.14)
\]
4.3. PENALIZED LIKELIHOOD RATIO TEST

or equivalently,

\[
\frac{1}{\sqrt{n}}p_{\ell_n'}(\theta) \approx N_5 \left( \sqrt{n}E_\theta \{ p_{\ell'}(Y_1, \theta) \} , Var_\theta \{ p_{\ell'}(Y_1, \theta) \} \right).
\]

Note that \( \sum_{i=1}^{n} \log f(y_i | \theta) \) is the log-likelihood function. Under the regularity conditions, \( E_\theta \{ \frac{\partial}{\partial \theta} \log f(Y_1 | \theta) \} = 0 \). Also, \( \frac{\partial}{\partial \theta} g(c, u_1) = \left( 0, 0, 0, 0, \{ \frac{u_1}{c} - \frac{(1 - u_1)}{(1 - c)} \} \right)^T \).

It is easy to see that \( E_\theta \{ p_{\ell'}(Y_1, \theta) \} = \left( 0, 0, 0, 0, \{ \frac{E(U_1)}{c} - \frac{(1 - E(U_1))}{(1 - c)} \} \right)^T \), where \( E(U_1) = 0.5 \).

For the maximum penalized likelihood estimator \( \hat{\theta} \), consider a Taylor series expansion

\[
p_{\ell_n'}(\hat{\theta}) = p_{\ell_n'}(\theta) + p_{\ell_n''}(\theta)(\hat{\theta} - \theta) + o_p(\sqrt{n}),
\]

where \( p_{\ell_n'}(\hat{\theta}) = \frac{\partial}{\partial \theta} p_{\ell_n}(\theta) \bigg|_{\theta = \hat{\theta}} = 0 \), and \( p_{\ell_n''}(\theta) = \frac{\partial^2}{\partial \theta \partial \theta} p_{\ell_n}(\theta) \).

It follows that

\[
\sqrt{n}(\hat{\theta} - \theta) = - \left\{ \frac{1}{n} p_{\ell_n''}(\theta) \right\}^{-1} \left\{ \frac{1}{\sqrt{n}} p_{\ell_n'}(\theta) \right\}.
\]

By Law of Large Number,

\[
\frac{1}{n} p_{\ell_n''}(\theta) = \frac{1}{n} \sum_{i=1}^{n} p_{\ell''}(Y_i, \theta) \xrightarrow{p} E_\theta \{ p_{\ell''}(Y_1, \theta) \} = -I_1(\theta),
\]

noting that \( E_\theta \{ p_{\ell''}(Y_1, \theta) \} = -I_1(\theta) \), under regularity condition C4.

Combining (4.14), (4.15), (4.16), and applying Slutsky’s theorem, the result i) follows for any \( \theta \in \Theta \) when \( \theta \) is the true parameter vector of model (4.2).
Next, consider the situation under the null hypothesis $H_0$, with $\beta_2 = \beta_3 = 0$, $c = \frac{1}{2}$. Then, $E_\theta \{p\ell'(Y_1, \theta^{(0)})\} = (0, 0, 0, 0, 0)^T$. Moreover,

$$Var_\theta \{p\ell'(Y_1, \theta^{(0)})\} = E_\theta \left[ p\ell'(Y_1, \theta^{(0)}) \left\{ p\ell'(Y_1, \theta^{(0)}) \right\}^T \right]$$

$$- \left[ E_\theta \{p\ell'(Y_1, \theta^{(0)})\} \right] \left[ E_\theta \{p\ell'(Y_1, \theta^{(0)})\} \right]^T = I_1(\theta^{(0)}).$$

Result $ii)$ follows from previous derivation, that is,

$$\sqrt{n} \left( \hat{\theta} - \theta^{(0)} \right) \overset{D}{\rightarrow} N_5 \left( 0, I_1^{-1}(\theta^{(0)}) \right),$$

under the null hypothesis $H_0$. \qed

Lemma 1 and Theorem 2 consider the properties of the penalized likelihood estimators of $\theta = (\beta, c)$ under the null hypothesis. Lemma 1 shows that $\hat{c}$ is bounded away from the parameter boundaries of $c = 0$ or $c = 1$, under the null hypothesis. Theorem 2 shows that, under the null hypothesis, $\hat{\beta}_2$ and $\hat{\beta}_3$ converge to their true values 0, in probability. In Theorem 3, we prove that the maximum penalized likelihood estimate of $\theta$ follows asymptotic multivariate normal distribution. These properties will be applied to derive the asymptotic null distribution of the penalized likelihood ratio test statistic in the next subsection.

### 4.3.4 Asymptotic null distribution of the penalized likelihood ratio test statistic $R_n$

We modify the classical proof of the asymptotic result for the likelihood ratio test statistic (Wilks (1938) [71]) for the penalized likelihood method.
Theorem 4. Assume that conditions C0- C5 hold. Consider the penalized log-likelihood ratio statistic $R_n$ defined in (4.4). Under $H_0$, the asymptotic distribution of $R_n$ is Chi-square distribution, that is,

$$R_n(\hat{\theta}) = 2\{p\ell_n(\hat{\theta}) - p\ell_n(\check{\theta})\} \overset{D}{\rightarrow} \chi^2_3,$$

(4.17)

as $n \rightarrow \infty$.

Proof. Apply Taylor series expansion to $p\ell_n(\theta)$ and use the fact that $p\ell_n'(\hat{\theta}) = 0$, we have

$$p\ell_n(\theta) = p\ell_n(\hat{\theta}) - \frac{1}{2}(\hat{\theta} - \theta)^T I(\hat{\theta})(\hat{\theta} - \theta) + o_p(1),$$

where

$$I(\hat{\theta}) = -\frac{\partial^2}{\partial \theta \partial \theta^T} p\ell_n(\theta)|_{\theta = \hat{\theta}} = -p\ell_n''(\hat{\theta}),$$

and $\hat{\theta}$ is the maximum penalized likelihood estimate on the entire parameter space $\Theta$. Let $pL_n(\theta)$ be the penalized likelihood function. Then,

$$pL_n(\theta) = \exp\{p\ell_n(\theta)\} = \exp\{p\ell_n(\hat{\theta})\} \exp\{-\frac{1}{2}(\hat{\theta} - \theta)^T I(\hat{\theta})(\hat{\theta} - \theta) + o_p(1)\}. \quad (4.18)$$

Now, consider (4.18) under $H_0$. We know that $\hat{\theta} \rightarrow \theta^{(0)}$ in probability by Theorem 2, and

$$\frac{1}{n} I(\theta^{(0)}) \overset{P}{\rightarrow} I_1(\theta^{(0)})$$

by Law of Large Numbers. Therefore,

$$\frac{1}{n} I(\hat{\theta}) \overset{P}{\rightarrow} I_1(\theta^{(0)}).$$
4.3. PENALIZED LIKELIHOOD RATIO TEST

By result \( ii \) of Theorem 3,

\[
\{nI_1(\theta)\}^{\frac{1}{2}} (\hat{\theta} - \theta) \overset{D}{\rightarrow} N_5(0, \mathbb{1}),
\]

where \( \mathbb{1} \) is a \( 5 \times 5 \) identity matrix, when \( \theta = \theta^{(0)} \). Let \( \Sigma = \{I_1(\theta)\}^{-1} \). Then,

\[
\frac{1}{n} I(\hat{\theta}) = \Sigma^{-1} + o_p(1) \text{ in (4.18)}.
\]

Write \( \theta = (\theta_1^T, \theta_2^T)^T \), where \( \theta_1 = (\beta_0, \beta_1)^T, \theta_2 = (\beta_2, \beta_3, c)^T \). Let \( z_1 = \sqrt{n} (\hat{\theta}_1 - \theta_1) \), \( z_2 = \sqrt{n} (\hat{\theta}_2 - \theta_2) \). Then, (4.18) can be written as

\[
pL_n(\theta) = \exp \{p\ell_n(\hat{\theta})\} \exp \left\{-\frac{1}{2} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}^T \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + o_p(1) \right\},
\]

(4.19)

where \[
\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \Sigma^{-1}, \text{ and } A_{11}, A_{12}, A_{21}, A_{22} \text{ are } 2 \times 2, 2 \times 3, 3 \times 2, 3 \times 3
\]

matrices. Expressing \( \Sigma \) as \[
\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}
\]

with \( 2 \times 2, 2 \times 3, 3 \times 2, 3 \times 3 \) matrices of \( \Sigma_{11}, \Sigma_{12}, \Sigma_{21}, \Sigma_{22} \), we notice block inversion matrix that

\[
A_{11} = (\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})^{-1}, \quad A_{12} = -A_{11} \Sigma_{12} \Sigma_{22}^{-1},
\]

\[
A_{22} = (\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12})^{-1}, \quad A_{21} = -A_{22} \Sigma_{21} \Sigma_{11}^{-1}.
\]

From (4.19), the log penalized likelihood can be written as

\[
p\ell_n(\theta) = p\ell_n(\hat{\theta}) - \frac{1}{2} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}^T \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + o_p(1). \tag{4.20}
\]
4.3. PENALIZED LIKELIHOOD RATIO TEST

We now show that (4.20) can be written as

\[
p_{\ell_n}(\theta) = p_{\ell_n}(\hat{\theta}) - \frac{1}{2} \left\{ (z_1 - \Sigma_{12} \Sigma_{22}^{-1} z_2)^T A_{11} (z_1 - \Sigma_{12} \Sigma_{22}^{-1} z_2) + z_2^T \Sigma_{22}^{-1} z_2 \right\} + o_p(1). \tag{4.21}
\]

In (4.20), consider the term

\[
\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}^T \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = z_1^T A_{11} z_1 + z_2^T A_{21} z_1 + z_1^T A_{12} z_2 + z_2^T A_{22} z_2. \tag{4.22}
\]

Let \( B \) be a 2 \times 3 matrix such that (4.22) can be written as

\[
\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}^T \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = (z_1 - B z_2)^T A_{11} (z_1 - B z_2) + z_2^T \Sigma_{22}^{-1} z_2, \tag{4.23}
\]

where \((z_1 - B z_2)\) is a linear combination of \(z_1, z_2\). The right hand side of (4.23) is

\[
z_1^T A_{11} z_1 - z_1^T A_{11} B z_2 - z_2^T B^T A_{11} z_1 + z_2^T B^T A_{11} B z_2 + z_2^T \Sigma_{22}^{-1} z_2.
\]

Comparing to the right hand side of (4.22), we need

\[
A_{12} = -A_{11} B, \quad A_{21} = -B^T A_{11}, \quad A_{22} = B^T A_{11} B + \Sigma_{22}^{-1}. \tag{4.24}
\]

It then requires that

\[
B = -A_{11}^{-1} A_{12} = A_{11}^{-1} (A_{11} \Sigma_{12} \Sigma_{22}^{-1}) = \Sigma_{12} \Sigma_{22}^{-1}.
\]
With this $B$ expression, all equations in (4.24) hold by noticing that $A_{21} = A_{12}^T$, and

$$B^T A_{11} B + \Sigma^{-1}_{22} = -A_{21} B + \Sigma^{-1}_{22} = A_{22} \Sigma_{21} \Sigma^{-1}_{11} \Sigma_{12} \Sigma^{-1}_{22} + \Sigma^{-1}_{22}$$

$$= A_{22}(\Sigma_{22} - A^{-1}_{22}) \Sigma^{-1}_{22} + \Sigma^{-1}_{22} = (A_{22} \Sigma_{22} - I) \Sigma^{-1}_{22} + \Sigma^{-1}_{22} = A_{22}.$$  

This confirms that (4.21) is an equivalent expression for $p\ell_n(\theta)$.

Next consider maximizing $p\ell_n(\theta)$ of the form (4.21) under the null hypothesis $H_0$, $p\ell_n(\theta)$ is now a function of $\theta_1 = (\beta_0, \beta_1)^T$ only. It is maximized when

$$z_1 - \Sigma_{12} \Sigma^{-1}_{22} z_{20} = 0,$$

i.e.

$$(\hat{\theta}_1 - \theta_1) - \Sigma_{12} \Sigma^{-1}_{22} (\hat{\theta}_2 - \theta_{20}) = 0,$$

where $z_{20} = \hat{\theta}_2 - \theta_{20}$, and $\theta_{20}$ is the true value of $\theta_2$ under $H_0$. The maximum penalized likelihood estimate under $H_0$ is

$$\hat{\theta}_1 = \hat{\theta}_1 - \Sigma_{12} \Sigma^{-1}_{22} (\hat{\theta}_2 - \theta_{20}),$$

and the maximum value of $p\ell_n(\theta)$ is

$$p\ell_n(\hat{\theta}_1, \theta_{20}) = p\ell_n(\hat{\theta}) - \frac{1}{2} z_{20}^T \Sigma^{-1}_{22} z_{20} + o_p(1).$$

It follows then under $H_0$,

$$R_n = 2 \left[ p\ell_n(\hat{\theta}) - p\ell_n(\hat{\theta}_1, \theta_{20}) \right] = z_{20}^T \Sigma^{-1}_{22} z_{20} + o_p(1).$$
4.4. SIMULATION STUDIES

By Theorems 2 and 3, we have
\[ R_n \xrightarrow{D} \chi^2_3. \]

\[ \square \]

4.3.5 Choice of penalty term

We can use any function of \( c \) which converges to infinity for \( c = 0, 1 \) as the penalty function. If we choose a penalty function that is a fixed function of \( c \), we have some problems when proving Theorems 3 and 4. More specifically,

\[ I_1(\theta) = E \left[ \{p\ell(Y_1, \theta)\}\{p\ell(Y_1, \theta)\}^T \right] \]

in Theorem 3 is not equal to \(-E\left[ \frac{\partial^2}{\partial \theta \partial \theta^T} p\ell(Y_1, \theta) \right]\) as for the regular likelihood approach, and \( \Sigma \) is not positive definite for \( \theta \in \Theta_0 \). These technical issues are resolved when we adopt the random penalty term \( g_n(c, u) \) that depends on the random variable \( U \sim Bernoulli(1, 0.5) \).

4.4 Simulation studies

4.4.1 Simulation model design

In this section, we conducted a set of numerical simulations to evaluate the finite sample properties of the proposed method. To check whether the penalized log-likelihood ratio test statistic follows an asymptotic \( \chi^2_3 \) distribution under the null hypothesis \( H_0 \), we run simulations with \( R = 2000 \) replications and reported the empirical test sizes. We further studied the empirical power of the test, under the alternative hypothesis \( H_1 \).
For different values of the regression coefficients, we replicated the simulation
$R = 2000$ times for data generated from model (4.1). We generated binary outcome
data from the Bernoulli distribution with

$$p = \frac{\exp(\beta_0 + \beta_1 z)}{1 + \exp(\beta_0 + \beta_1 z)},$$

under the null hypothesis, $H_0 : \beta_2 = \beta_3 = 0$, to evaluate the size of the test. Also,
the treatment was randomly assigned to 50% of the patients, with $P(Z = 1) = 0.5$.
The biomarker values, $x$, were generated from a uniform distribution on the interval
$(0,1)$.

In all the numerical studies below, we used model (4.2) as an approximation to
model (4.1). In order to calculate the value of the penalized likelihood ratio test
statistic $R_n$, we need to obtain the maximum penalized likelihood estimate of the
parameter $\theta$ on the parameter space $\Theta$. To get $\hat{\theta}$, we used Newton-Raphson method
in which

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + V^{-1}(\hat{\theta}^{(k)})S(\hat{\theta}^{(k)}),$$

where $S$ is the score function containing the first-order derivatives of the penalized
likelihood function with respect to $\theta$,

$$S = \frac{\partial}{\partial \theta} p\ell_n(\theta) = (S(\beta_0), S(\beta_1), S(\beta_2), S(\beta_3), S(c))^T,$$

where $S(\beta_0) = \sum_{i=1}^n (y_i - p_i)$, $S(\beta_1) = \sum_{i=1}^n z_i(y_i - p_i)$, $S(\beta_2) = \sum_{i=1}^n w(x_i, c)(y_i - p_i)$,
$S(\beta_3) = \sum_{i=1}^n z_iw(x_i, c)(y_i - p_i)$, and $S(c) = \sum_{i=1}^n -Kw(x_i, c)(1 - w(x_i, c))(y_i - p_i)(\beta_2 + \beta_3 z_i) + \left(\frac{U}{c} - \frac{n-U}{1-c}\right)$, in which $U \sim Binomial(n, 0.5)$. Also, $V$ is a $5 \times 5$ matrix
containing the negative second-order derivatives of the penalized likelihood function.
with respect to $\theta$, where $v_{ij} = -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell_n(\theta)$, for $i, j = 1, \ldots, 5$ in which

\[
v_{11} = \sum_{i=1}^{n} p_i(1 - p_i), \quad v_{12} = v_{21} = \sum_{i=1}^{n} z_i p_i(1 - p_i),
\]

\[
v_{13} = v_{31} = \sum_{i=1}^{n} w(x_i, c)p_i(1 - p_i), \quad v_{14} = v_{41} = \sum_{i=1}^{n} z_i w(x_i, c)p_i(1 - p_i),
\]

\[
v_{15} = v_{51} = \sum_{i=1}^{n} -K(\beta_2 + \beta_3 z_i)w(x_i, c)(1 - w(x_i, c))p_i(1 - p_i),
\]

\[
v_{22} = \sum_{i=1}^{n} z_i^2 p_i(1 - p_i), \quad v_{23} = v_{32} = \sum_{i=1}^{n} z_i w(x_i, c)p_i(1 - p_i),
\]

\[
v_{24} = v_{42} = \sum_{i=1}^{n} z_i^2 w(x_i, c)p_i(1 - p_i),
\]

\[
v_{25} = v_{52} = \sum_{i=1}^{n} -Kz_i(\beta_2 + \beta_3 z_i)w(x_i, c)(1 - w(x_i, c))p_i(1 - p_i),
\]

\[
v_{33} = \sum_{i=1}^{n} w^2(x_i, c)p_i(1 - p_i), \quad v_{34} = v_{43} = \sum_{i=1}^{n} z_i w^2(x_i, c)p_i(1 - p_i),
\]

\[
v_{35} = v_{53} = \sum_{i=1}^{n} Kw(x_i, c)(1 - w(x_i, c))[(y_i - p_i) - \{(\beta_2 + \beta_3 z_i)w(x_i, c)p_i(1 - p_i)\}],
\]

\[
v_{44} = \sum_{i=1}^{n} w^2(x_i, c)z_i^2 p_i(1 - p_i),
\]

\[
v_{45} = v_{54} = \sum_{i=1}^{n} Kw(x_i, c)(1 - w(x_i, c))[z_i(y_i - p_i) - \{(\beta_2 + \beta_3 z_i)w(x_i, c)z_i p_i(1 - p_i)\}],
\]

\[
v_{55} = \sum_{i=1}^{n} Kw(x_i, c)(1 - w(x_i, c))(\beta_2 + \beta_3 z_i) [K(2w(x_i, c) - 1)(y_i - p_i)
\]

\[+K\{w(x_i, c)(1 - w(x_i, c))(\beta_2 + \beta_3 z_i)p_i(1 - p_i)\} + \frac{U}{c^2} + \frac{n - U}{(1 - c)^2}.\]
We used the profile penalized likelihood estimate of $\theta$ as an initial value for the Newton-Raphson algorithm. Using this initial value, we can update the value of $\theta$ until convergence, i.e., the algorithm will be continued until

$$||\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}|| = \max\{|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}|\} < \epsilon.$$ 

We used $\epsilon = 0.001$ in our simulations.

Note that the maximum penalized likelihood estimate of $\theta$ on the parameter space $\Theta_0$ is exactly the maximum likelihood estimate of $\theta$.

We did simulations for different values of $n$, $\beta_0$, and $\beta_1$. We considered sample sizes $n = 500,900$, and in $w(x,c)$ we took $K = 10,50,100,200,300,500$. Also, different combinations of parameters $\beta_0$ and $\beta_1 = \log(0.1), \log(0.5), \log(1.0), \log(1.5), \log(2.0), \log(3.0)$ were used to run simulations in order to evaluate empirical test sizes.

To evaluate the empirical power of the test, data must be generated under the alternative hypothesis, $H_1: \text{at least one of } \beta_2 \text{ or } \beta_3 \text{ is not equal to 0}$, with

$$p = \frac{\exp\{\beta_0 + \beta_1 z + \beta_2 I(x > c) + \beta_3 z I(x > c)\}}{1 + \exp\{\beta_0 + \beta_1 z + \beta_2 I(x > c) + \beta_3 z I(x > c)\}},$$

in model (4.1). We calculated the empirical power for different values of parameters $\beta_2$ and $\beta_3 = \log(0.5), \log(1.5), \log(2.0)$ in our simulations. Also, $c = 0.5, 0.75$ was used to generate the data set under the alternative hypothesis.
4.4. SIMULATION STUDIES

4.4.2 Simulations results

Empirical test sizes for different parameter combinations are presented in Tables 4.1, 4.2. The empirical size of the test is obtained as the proportion of replicated simulations that reached the pre-specified nominal significance level, for data generated under the null hypothesis. If the empirical test size is close to the significance level \( \alpha = 0.05 \), we can say the proposed method works well for testing the null hypothesis.

As we can see in Tables 4.1, 4.2, the empirical test sizes for different values of the parameter combination are close to the nominated significance level \( \alpha = 0.05 \). This means that the proposed penalized likelihood ratio test statistic, following asymptotic chi-square distribution, is accurate in terms of type I error for testing \( H_0 : \beta_2 = \beta_3 = 0 \).
4.4. SIMULATION STUDIES

Table 4.1: Empirical test sizes for testing $H_0 : \beta_2 = \beta_3 = 0^1$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$K$</th>
<th>$\exp(\beta_0)$</th>
<th>$\exp(\beta_1)$</th>
<th>Rejection Rate(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>10</td>
<td>0.1</td>
<td>1.0</td>
<td>4.9</td>
</tr>
<tr>
<td>500</td>
<td>10</td>
<td>0.5</td>
<td>1.0</td>
<td>4.7</td>
</tr>
<tr>
<td>500</td>
<td>10</td>
<td>0.5</td>
<td>1.5</td>
<td>4.3</td>
</tr>
<tr>
<td>500</td>
<td>10</td>
<td>2.0</td>
<td>2.0</td>
<td>4.6</td>
</tr>
<tr>
<td>500</td>
<td>10</td>
<td>2.0</td>
<td>3.0</td>
<td>3.8</td>
</tr>
<tr>
<td>900</td>
<td>10</td>
<td>0.1</td>
<td>1.0</td>
<td>5.3</td>
</tr>
<tr>
<td>900</td>
<td>10</td>
<td>0.5</td>
<td>1.0</td>
<td>5.8</td>
</tr>
<tr>
<td>900</td>
<td>10</td>
<td>0.5</td>
<td>1.5</td>
<td>6.0</td>
</tr>
<tr>
<td>900</td>
<td>10</td>
<td>2.0</td>
<td>2.0</td>
<td>5.4</td>
</tr>
<tr>
<td>900</td>
<td>10</td>
<td>2.0</td>
<td>3.0</td>
<td>5.9</td>
</tr>
<tr>
<td>500</td>
<td>25</td>
<td>0.1</td>
<td>1.0</td>
<td>5.5</td>
</tr>
<tr>
<td>500</td>
<td>25</td>
<td>0.5</td>
<td>1.0</td>
<td>5.5</td>
</tr>
<tr>
<td>500</td>
<td>25</td>
<td>0.5</td>
<td>1.5</td>
<td>5.5</td>
</tr>
<tr>
<td>500</td>
<td>25</td>
<td>2.0</td>
<td>2.0</td>
<td>5.3</td>
</tr>
<tr>
<td>500</td>
<td>25</td>
<td>2.0</td>
<td>3.0</td>
<td>5.3</td>
</tr>
<tr>
<td>900</td>
<td>25</td>
<td>0.1</td>
<td>1.0</td>
<td>5.3</td>
</tr>
<tr>
<td>900</td>
<td>25</td>
<td>0.5</td>
<td>1.0</td>
<td>5.2</td>
</tr>
<tr>
<td>900</td>
<td>25</td>
<td>0.5</td>
<td>1.5</td>
<td>5.1</td>
</tr>
<tr>
<td>900</td>
<td>25</td>
<td>2.0</td>
<td>2.0</td>
<td>5.6</td>
</tr>
<tr>
<td>900</td>
<td>25</td>
<td>2.0</td>
<td>3.0</td>
<td>5.2</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>0.1</td>
<td>1.0</td>
<td>4.7</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>0.5</td>
<td>1.0</td>
<td>5.4</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>0.5</td>
<td>1.5</td>
<td>4.6</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>2.0</td>
<td>2.0</td>
<td>4.5</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>2.0</td>
<td>3.0</td>
<td>4.1</td>
</tr>
<tr>
<td>900</td>
<td>50</td>
<td>0.1</td>
<td>1.0</td>
<td>5.7</td>
</tr>
<tr>
<td>900</td>
<td>50</td>
<td>0.5</td>
<td>1.0</td>
<td>5.9</td>
</tr>
<tr>
<td>900</td>
<td>50</td>
<td>0.5</td>
<td>1.5</td>
<td>5.6</td>
</tr>
<tr>
<td>900</td>
<td>50</td>
<td>2.0</td>
<td>2.0</td>
<td>5.3</td>
</tr>
<tr>
<td>900</td>
<td>50</td>
<td>2.0</td>
<td>3.0</td>
<td>5.7</td>
</tr>
</tbody>
</table>

$^1$ Significance level $\alpha = 0.05$. Binary data were generated based on the logistic regression model (4.1), under the null hypothesis. Biomarker values $\sim U(0,1)$. The new treatment and the control arm were assigned randomly to the patients, and $R = 2000$ replications.
4.4. SIMULATION STUDIES

Table 4.2: Empirical test sizes for testing \( H_0 : \beta_2 = \beta_3 = 0 \).

<table>
<thead>
<tr>
<th>( n )</th>
<th>( K )</th>
<th>( n \exp(\beta_0) )</th>
<th>( \exp(\beta_1) )</th>
<th>Rejection Rate(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>100</td>
<td>0.1</td>
<td>1.0</td>
<td>5.0</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>0.5</td>
<td>1.0</td>
<td>5.6</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>0.5</td>
<td>1.5</td>
<td>5.3</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>2.0</td>
<td>2.0</td>
<td>4.7</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>2.0</td>
<td>3.0</td>
<td>4.4</td>
</tr>
<tr>
<td>900</td>
<td>100</td>
<td>0.1</td>
<td>1.0</td>
<td>5.8</td>
</tr>
<tr>
<td>900</td>
<td>100</td>
<td>0.5</td>
<td>1.0</td>
<td>5.7</td>
</tr>
<tr>
<td>900</td>
<td>100</td>
<td>0.5</td>
<td>1.5</td>
<td>5.3</td>
</tr>
<tr>
<td>900</td>
<td>100</td>
<td>2.0</td>
<td>2.0</td>
<td>5.3</td>
</tr>
<tr>
<td>900</td>
<td>100</td>
<td>2.0</td>
<td>3.0</td>
<td>5.7</td>
</tr>
<tr>
<td>500</td>
<td>200</td>
<td>0.1</td>
<td>1.0</td>
<td>5.4</td>
</tr>
<tr>
<td>500</td>
<td>200</td>
<td>0.5</td>
<td>1.0</td>
<td>6.1</td>
</tr>
<tr>
<td>500</td>
<td>200</td>
<td>0.5</td>
<td>1.5</td>
<td>5.4</td>
</tr>
<tr>
<td>500</td>
<td>200</td>
<td>2.0</td>
<td>2.0</td>
<td>4.9</td>
</tr>
<tr>
<td>500</td>
<td>200</td>
<td>2.0</td>
<td>3.0</td>
<td>5.7</td>
</tr>
<tr>
<td>900</td>
<td>200</td>
<td>0.1</td>
<td>1.0</td>
<td>6.1</td>
</tr>
<tr>
<td>900</td>
<td>200</td>
<td>0.5</td>
<td>1.0</td>
<td>6.0</td>
</tr>
<tr>
<td>900</td>
<td>200</td>
<td>0.5</td>
<td>1.5</td>
<td>5.8</td>
</tr>
<tr>
<td>900</td>
<td>200</td>
<td>2.0</td>
<td>2.0</td>
<td>5.2</td>
</tr>
<tr>
<td>900</td>
<td>200</td>
<td>2.0</td>
<td>3.0</td>
<td>5.7</td>
</tr>
<tr>
<td>500</td>
<td>300</td>
<td>0.1</td>
<td>1.0</td>
<td>6.4</td>
</tr>
<tr>
<td>500</td>
<td>300</td>
<td>0.5</td>
<td>1.0</td>
<td>5.7</td>
</tr>
<tr>
<td>500</td>
<td>300</td>
<td>0.5</td>
<td>1.5</td>
<td>5.4</td>
</tr>
<tr>
<td>500</td>
<td>300</td>
<td>2.0</td>
<td>2.0</td>
<td>6.1</td>
</tr>
<tr>
<td>500</td>
<td>300</td>
<td>2.0</td>
<td>3.0</td>
<td>5.8</td>
</tr>
<tr>
<td>900</td>
<td>300</td>
<td>0.1</td>
<td>1.0</td>
<td>5.2</td>
</tr>
<tr>
<td>900</td>
<td>300</td>
<td>0.5</td>
<td>1.0</td>
<td>5.7</td>
</tr>
<tr>
<td>900</td>
<td>300</td>
<td>0.5</td>
<td>1.5</td>
<td>6.0</td>
</tr>
<tr>
<td>900</td>
<td>300</td>
<td>2.0</td>
<td>2.0</td>
<td>5.9</td>
</tr>
<tr>
<td>900</td>
<td>300</td>
<td>2.0</td>
<td>3.0</td>
<td>6.1</td>
</tr>
</tbody>
</table>

\(^1\) Significance level \( \alpha = 0.05 \). Binary data were generated based on the logistic regression model (4.1), under the null hypothesis. Biomarker values \( \sim U(0,1) \). The new treatment and the control arm were assigned randomly to the patients, and \( R = 2000 \) replications.

To further study the empirical distribution of the test statistic under \( H_0 : \beta_2, \beta_3 = \)
0, we calculated the quantiles for the penalized log-likelihood ratio test statistic for a simulated data set with $n = 900$, $K = 100$, $\beta_0 = \log(0.5)$, and $\beta_1 = \log(1.5)$. These quantiles were plotted with quantiles of $\chi^2_3$ in Figure 4.2.

![Figure 4.2: Q-Q plot of observed values of $R_n$ and $\chi^2_3$](image)

As we can see in Figure 4.2, the values in the Q-Q plot are close to the straight line. This indicates that the penalized likelihood ratio test statistic has an asymptotic distribution of Chi-square distribution with 3 degrees of freedom.

We evaluated the empirical power for different non-zero values of $\beta_2$ or $\beta_3$ and reported the results in Table 4.3. To calculate the power, data were generated under the alternative hypothesis. The empirical power is the percentage of the replicated simulations that reached the pre-specified significance level, when data are generated
under the alternative hypothesis.

Table 4.3: Empirical power of the test under $H_1$ : at least one of $\beta_2$ or $\beta_3$ is not equal to 0, with $c = 0.5^1$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$K$</th>
<th>$\text{exp}(\beta_0)$</th>
<th>$\text{exp}(\beta_1)$</th>
<th>$\text{exp}(\beta_2)$</th>
<th>$\text{exp}(\beta_3)$</th>
<th>Rejection Rate(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>10</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>15.3</td>
</tr>
<tr>
<td>500</td>
<td>10</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>54.9</td>
</tr>
<tr>
<td>500</td>
<td>10</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>93.6</td>
</tr>
<tr>
<td>900</td>
<td>10</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>24.0</td>
</tr>
<tr>
<td>900</td>
<td>10</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>80.1</td>
</tr>
<tr>
<td>900</td>
<td>10</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>99.9</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>16.8</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>59.6</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>96.0</td>
</tr>
<tr>
<td>900</td>
<td>50</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>28.4</td>
</tr>
<tr>
<td>900</td>
<td>50</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>84.5</td>
</tr>
<tr>
<td>900</td>
<td>50</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>100</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>17.6</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>65.1</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>2.5</td>
<td>84.8</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>96.0</td>
</tr>
<tr>
<td>900</td>
<td>100</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>28.8</td>
</tr>
<tr>
<td>900</td>
<td>100</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>85.2</td>
</tr>
<tr>
<td>900</td>
<td>100</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>100</td>
</tr>
<tr>
<td>500</td>
<td>200</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>18.9</td>
</tr>
<tr>
<td>500</td>
<td>200</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>58.3</td>
</tr>
<tr>
<td>500</td>
<td>200</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>97.5</td>
</tr>
<tr>
<td>900</td>
<td>200</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>28.5</td>
</tr>
<tr>
<td>900</td>
<td>200</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>63.4</td>
</tr>
<tr>
<td>900</td>
<td>200</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>100</td>
</tr>
</tbody>
</table>

$^1$Significance level $\alpha = 0.05$. Binary data were generated based on the logistic regression model (4.1), with $c = 0.5$, under the alternative hypothesis. Biomarker values follows $U(0, 1)$. The new treatment and the control arm were assigned randomly to the patients, and $R = 2000$ replications.
4.4. SIMULATION STUDIES

Table 4.4: Empirical power of the test under $H_1$: at least one of $\beta_2$ or $\beta_3$ is not equal to 0, with $c = 0.75^1$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$K$</th>
<th>$\exp(\beta_0)$</th>
<th>$\exp(\beta_1)$</th>
<th>$\exp(\beta_2)$</th>
<th>$\exp(\beta_3)$</th>
<th>Rejection Rate(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>50</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>16.7</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>33.9</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>2.0</td>
<td>2.0</td>
<td>1.0</td>
<td>3.0</td>
<td>19.3</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>2.0</td>
<td>2.0</td>
<td>1.0</td>
<td>4.0</td>
<td>24.5</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>40.2</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>3.0</td>
<td>59.0</td>
</tr>
<tr>
<td>900</td>
<td>50</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>30.7</td>
</tr>
<tr>
<td>900</td>
<td>50</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>62.0</td>
</tr>
<tr>
<td>900</td>
<td>50</td>
<td>2.0</td>
<td>2.0</td>
<td>1.0</td>
<td>3.0</td>
<td>30.6</td>
</tr>
<tr>
<td>900</td>
<td>50</td>
<td>2.0</td>
<td>2.0</td>
<td>1.0</td>
<td>4.0</td>
<td>39.3</td>
</tr>
<tr>
<td>900</td>
<td>50</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>71.3</td>
</tr>
<tr>
<td>900</td>
<td>50</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>3.0</td>
<td>88.7</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>17.4</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>33.0</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>2.0</td>
<td>2.0</td>
<td>1.0</td>
<td>3.0</td>
<td>19.5</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>2.0</td>
<td>2.0</td>
<td>1.0</td>
<td>4.0</td>
<td>24.7</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>43.3</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>3.0</td>
<td>60.6</td>
</tr>
<tr>
<td>900</td>
<td>100</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>29.9</td>
</tr>
<tr>
<td>900</td>
<td>100</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>60.5</td>
</tr>
<tr>
<td>900</td>
<td>100</td>
<td>2.0</td>
<td>2.0</td>
<td>1.0</td>
<td>3.0</td>
<td>29.3</td>
</tr>
<tr>
<td>900</td>
<td>100</td>
<td>2.0</td>
<td>2.0</td>
<td>1.0</td>
<td>4.0</td>
<td>38.8</td>
</tr>
<tr>
<td>900</td>
<td>100</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>69.7</td>
</tr>
<tr>
<td>900</td>
<td>100</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>3.0</td>
<td>86.6</td>
</tr>
</tbody>
</table>

$^1$ Significance level $\alpha = 0.05$. Binary data were generated based on the logistic regression model (4.1), with $c = 0.75$, under the alternative hypothesis. Biomarker values follows $U(0, 1)$. The new treatment and the control arm were assigned randomly to the patients, and $R = 2000$ replications.

As we can see in the Table 4.3 (with $c = 0.5$), the penalized likelihood ratio test works well, and the empirical power of the test is close to 1 or equal to 1 in some cases. Also, we calculated the power of the test for the case that $\beta_2 = 0$, and $\beta_3 \neq 0$ because the more interesting parameter is $\beta_3$. As we can see in Table 4.4 (with $c = 0.75$), the power is low when the value of $\beta_3$ is small. This happens because for $c = 0.75$, only
25% of patients benefit more from the new treatment. Therefore, the small values of
the parameter are detected with lower power. However, for the larger values of $\beta_3$ or
larger sample sizes, the test becomes more powerful.

For sample size $n = 900$, $K = 100$ with $c = 0.5$, $\beta_2 = \log(0.5)$, $\beta_3 = \log(1.5)$ will
ensure the study has more 85% power to detect a subset effect. However, for $n = 500$,
we need $\beta_3 = \log(2.5)$ to have 85% power.

![Diagram](image)

(a) using $I(x > c)$  
(b) using $w(x, c)$

Figure 4.3: $p\ell_n(\hat{\beta}_c, c)$ vs. $c$. Data were generated under $H_0$.

Under $H_0$, to compare how functions $w(x, c)$ and $I(x > c)$ affect estimating the
unknown cut-point $c$, we plotted the penalized likelihood function vs. different values
of $c$ using $w(x, c)$ to estimate $c$ by Newton-Raphson method and $I(x > c)$ to estimate
$c$ by profile penalized likelihood method. For the profile penalized likelihood method,
we used a grid search of $c$ values between 0.01, 0.99, increased by 0.01. Figure 4.3(a)
presents the penalized likelihood function, $p\ell_n(\hat{\beta}_c, c)$, vs. different values of $0 < c < 1$
when data are generated under the null hypothesis with $\beta_0 = \log(0.5)$, $\beta_1 = \log(1.5)$,
$n = 900$, $K = 100$ in model (4.1), using $I(x > c)$. The vertical line identifies the
value of $c$ that maximizes the penalized likelihood function. Using $I(x > c)$ and
the profile penalized likelihood method, \( \hat{c}_P = 0.51 \). Also, Figure 4.3(b) presents the penalized likelihood function vs. different values of \( 0 < c < 1 \) when data are generated under the null hypothesis with above setting but fitted by model (4.2), using \( w(x, c) \). Using \( w(x, c) \) and Newton-Raphson method, we obtain \( \hat{c}_N = 0.5105 \), which is the maximum penalized likelihood estimate of \( c \). The plot shows that the profile penalized likelihood estimate of unknown cut-point \( c \) is close to the maximum penalized likelihood estimate of \( c \). It confirms that \( w(x, c) \) is a good approximation for \( I(x > c) \). It also confirms Theorem 2 that the maximum penalized likelihood estimate of \( c \) converges to 0.5, under \( H_0 \).

The estimation of \( c \) under \( H_1 \) should be based on the likelihood function, not the penalized likelihood function. Having a penalty term in the penalized likelihood function may introduce bias in the estimation of \( c \), under \( H_1 \). However, we use the penalized likelihood method for hypothesis testing mainly, and will explore the estimation of \( \beta \) and \( c \) in the future.

### 4.5 Application to prostate cancer data set

To demonstrate the use of the proposed penalized likelihood method for testing the main biomarker effect and the treatment-biomarker interaction effect, we applied the method to data from Veterans Administration Cooperative Urologic Research Group clinical trial (Byar and Corle (1977), Andrews and Herzberg (1985)[10, 5]). This double-blind clinical trial randomly allocated 505 prostate cancer patients to one of four levels: placebo, 0·2 mg of diethylstilbestrol (DES), 1·0 mg DES, 5·0 mg DES. In our analysis, the two lower doses were combined as control group, and the two higher doses were combined as treatment group. The treatment is assigned to 74·6 % of
the patients. We used serum Prostatic Acid Phosphatase level (AP) as a biomarker that is transformed to uniform scale on $(0,1)$ in the analysis using percentile of the original measurement. The original data have survival outcomes, and we consider instead 3-year survival (36 months) as outcome of interest. All patients who died before 36 months with $T_i < 36$ will have $Y_i = 1$ (censored observations are treated as missing values. The number of missing values was 0 because none of the patients with $T_i < 36$ were censored). Patients who were alive at 36 months will have $Y_i = 0$. The number of observed deaths before 36 months is 257. Also, the number of patients that are alive at 36 months is 248.

To test the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$, we applied the proposed method to data. We obtained $p$-value=0.0105, based on a $\chi^2_3$ asymptotic distribution for the penalized likelihood ratio test statistic, for testing $H_0$, which is highly significant for significant level $\alpha = 0.05$. Therefore, the null hypothesis is rejected, i.e., there is either a significant biomarker main effect or a significant biomarker-treatment interaction effect. This implies a biomarker cut-point $c$ exists and the next step is to examine if there is a significant treatment-biomarker interaction effect.

Now that $H_0$ is rejected, the estimation should be based on the likelihood function for model (4.2) (without the penalty term). Notice that under $H_1$, model (4.2) is a regular model and identifiable in terms of the parameters. The maximum likelihood estimate of $c$ by maximizing the likelihood function is $\hat{c} = 0.774$ on the transformed scale, which is equivalent to $\hat{c} = 38.00$ in the original scale of AP.
4.6 Discussions

In this chapter, we developed a penalized likelihood ratio test for testing the biomarker main effect and the treatment-biomarker interaction effect for binary outcome variable when the biomarker cut-point is unknown. The challenges are from three aspects of the biomarker threshold models, 1) the indicate function for the threshold effect is not a continuous function with respect to the threshold parameter $c$, 2) under the null hypothesis, the biomarker threshold model is not identifiable with respect to different values of $c$ when $\beta_2 = \beta_3 = 0$, and 3) the parameter values of $c$ lie on the boundary of $[0, 1]$ under $H_0$.

We approximated the non-continuous indicator function $I(x > c)$ by the continuous function $w(x, c)$. Then, we showed that using penalty term overcomes irregularities of model (4.1) and allow us to derive a simple limiting distribution for the test statistic under the null hypothesis. Through extensive simulation studies, we confirmed that the asymptotic result developed for model (4.2) gives accurate empirical test size when the sample size is large ($n \geq 500$). For small values of $\beta_2$ and $\beta_3$, the empirical power of the test is small, but the power increases quickly as the parameter values are further away from the null values. Overall, the proposed test method provides reasonable statistical power under the alternative hypothesis for moderate to large effects of parameters $\beta_2$ and $\beta_3$.

The penalized likelihood ratio test for model (4.2) provides a simple and neat approach to detect the biomarker main effect and the treatment-biomarker interaction in model (4.1). The introduction of the random penalty term is a novel idea to deal with the non-smooth and irregular model (4.1), and it leads to a test statistic that has a very simple asymptotic distribution. Unlike the usual approach to the penalized
likelihood method, which requires the selection of tuning parameter, our approach of random penalty term does not have this kind of restriction. The use of the penalty term is for conducting a test of $H_0$ only. When $H_0$ is rejected, then estimation of $\beta$ and $c$ should be based on the likelihood function for model (4.2). Some computational methods were developed for testing the treatment-biomarker interaction. Chen et al. (2014) [11] applied a hierarchical Bayesian method, and Jiang et al. (2007) [39] used a biomarker-adaptive threshold design for testing the treatment-biomarker interaction effect. These previous methods deal with survival outcome data but can be extended to binary outcome data easily. The estimation of $\beta$ and $c$ will be further investigated in the future for model (4.2).

We introduced the penalized likelihood ratio test in the framework of the logistic regression model for binary outcome data. The method is in fact valid for any model with a likelihood function satisfying the conditions C0-C5. For example, the method works for generalized linear models, where the distributions of responses are in the exponential family.

Further extensions to survival outcome data will be considered in the future. The advantage of a simple asymptotic test method is evident when compared to the bootstrap methods, in terms of computation time. For example, each replication for simulations in chapter 2 takes 5 minutes to run, and for simulations in chapter 3 takes 25 minutes to run. However, for simulations based on asymptotic results of penalized likelihood method in chapter 4, it only takes 5 seconds to run one replication.
Chapter 5

Summary and conclusions

In this chapter, we summarize what we did in this thesis. Also, we will give some directions about the interesting topics for future work.

5.1 Summary

The importance of studying the treatment-biomarker interaction effect in biomarker threshold models motivated us to develop resampling methods and theoretical approaches in statistical inference, and we focused on the test methods. Existence of the interaction between the treatment and the biomarker means that a subset of patients benefit more from the new treatment. Identifying this subset of patients (by a predictive biomarker and a cut-point value) can help us to avoid unnecessary therapies and making optimal treatment decisions.

The big challenge is the existence of an unknown biomarker cut-point in many situations that causes irregularities and non-identifiability problems in the analysis of the statistical models. In this thesis, we tested for the treatment-biomarker interaction effects by porposing a residual bootstrap method or developing new asymptotic approaches, such as penalized likelihood method, to get over the irregularities and
non-identifiablity problem. Unlike available research in applying penalized likelihood method to conventional mixture models, we introduced a new idea of using a random penalty term in order to establish the limiting distribution of the proposed test statistic theoretically. The proposed Binary random penalty form can also be replaced by other penalty forms that meet regularity conditions. Moreover, we considered both survival outcome data (independent and clustered observations) in chapters 2 and 3 and binary outcome data in chapter 4.

5.2 Future work

There are some research questions that we are interested in finding the answers to. However, we could not work on them for this thesis because of time limitation.

Considering the proposed residual bootstrap test for clustered outcome data in chapter 3, the random effect is added to the random intercept. As future subject of interest, we can consider a random slope model. Therefore, the proposed residual bootstrap test can be extended for this case.

A question remained in chapter 4 is estimating the parameters $\beta$ and $c$ when the null hypothesis is rejected. In this situation, the regularity conditions hold under the alternative hypothesis for model (4.2). Therefore, there is no need to use the penalized likelihood function, and the model parameters can be estimated by maximizing the likelihood function. Rejection of $H_0 : \beta_2 = \beta_3 = 0$ means that either $\beta_2$ or $\beta_3$ is not equal to 0, which indicates that a biomarker cut-point $c$ exists. Further estimation and inference about $\beta_3$, the treatment-biomarker interaction effect, can be based on model (4.2) directly. Note that model (4.2) is a regular model under the alternative hypothesis.
Another important aspect of future research is to establish the penalized likelihood method for survival outcome data, because time to event is often the variable of interest in clinical trials. In this situation, the proportional hazards model and accelerated failure time model (AFT) are frequently used to fit a model for time to a clinical event outcome and its association with explanatory variables (Cox (1972) [21]).

We plan to extend the penalized likelihood method with random penalty to the Cox model and derive the limiting distribution of the proposed penalized log-likelihood ratio test statistic. Based on our results from the logistic regression model for binary outcome data in chapter 4, it is logical to expect that the penalized likelihood ratio statistic has a limiting Chi-square distribution under the null hypothesis.

Accelerate failure time (AFT) or log location-scale model is one of the most important models in survival analysis. An accelerate failure time model for response variable $T$, and covariates $Z = (Z_1, ..., Z_p)$ can be defined in the following way. Consider $Y = \log(T)$ and let $W$ be a random variable with probability density function, $g(w)$. Suppose $Y$ can be expressed as

$$Y = \mu(Z) + \sigma W,$$

and its density function has the form,

$$f(y|z) = \frac{1}{\sigma} g\left(\frac{y - \mu(z)}{\sigma}\right),$$

where $\mu(z) = \beta_0 + \beta_1 Z_1 + ... + \beta_p Z_p$; $\beta = (\beta_0, ..., \beta_p)^T$ is the vector of the regression coefficients. The failure time $T$ is then said to have an AFT model. It can be shown
that for an AFT model,

\[ S_T(t) = S_0(e^{-\mu(z)t}), \]

where \( S_0 \) is the baseline survival function (Lawless (2002) [44]). Unknown parameters can be estimated by the maximum likelihood method. For \( t_i, \delta_i \) defined in Section 1.2.1, the likelihood function has the form

\[ L(\beta, \sigma) = \prod_{i=1}^{n} \left\{ f(t_i) \right\}^{\delta_i} \left\{ S(t_i) \right\}^{1-\delta_i}. \]

Suppose we want to use an AFT model to associate the response variable time \( T \) to treatment \( Z \) and the biomarker \( X \). The model takes the form

\[ Y = \log(T) = \beta_0 + \beta_1 z + \beta_2 I(x > c) + \beta_3 z I(x > c) + \sigma W, \quad (5.1) \]

where the cut-point value \( c \) is unknown.

To extend the method of chapter 4 to AFT model, we will consider replacing \( I(x > c) \) by the smooth approximation \( w(x, c) \) in model (5.1), and also consider the random penalty approach. We expect that the penalized likelihood ratio statistic based on an AFT model has a limiting Chi-square distribution under the null hypothesis as well.

The censoring issue makes such extension to survival outcome data challenging, as the likelihood functions for the Cox model and AFT model do not have the same structure as that for the logistic regression model in chapter 4, or for a generalized linear model in general.
Bibliography


[22] R.B. Davies. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika, 64 , 247-254, 1977.*


[37] PhD Hilsenbeck, S.G. Prognostic and predictive markers: Whats the difference? & why should i care? Division of Biostatistics Smith Breast Center and Duncan Cancer Center Baylor College of Medicine Houston, TX, 2014.


