6-11-2010

Semi-Empirical Likelihood Confidence Intervals for the ROC Curve with Missing Data

Xiaoxia Liu

SEMI-EMPIRICAL LIKELIHOOD CONFIDENCE INTERVALS FOR THE ROC CURVE WITH MISSING DATA

by

XIAOXIA LIU

Under the Direction of Dr. Yichuan Zhao

ABSTRACT

The receiver operating characteristic (ROC) curve is one of the most commonly used methods to compare the diagnostic performances of two or more laboratory or diagnostic tests. In this thesis, we propose semi-empirical likelihood based confidence intervals for ROC curves of two populations, where one population is parametric while the other one is non-parametric and both populations have missing data. After imputing missing values, we derive the semi-empirical likelihood ratio statistic and the corresponding likelihood equations. It has been shown that the log-semi-empirical likelihood ratio statistic is asymptotically chi-square distributed. The estimating equations are solved simultaneously to obtain the estimated lower and upper bounds of semi-empirical likelihood confidence intervals. Simulation studies are conducted to evaluate the finite sample performance of the proposed empirical likelihood confidence intervals with various sample sizes and different missing rates.

INDEX WORDS: Confidence interval, Empirical likelihood, Estimating equation, Missing data, ROC curve

SEMI-EMPIRICAL LIKELIHOOD CONFIDENCE INTERVALS FOR THE ROC CURVE WITH MISSING DATA

by

XIAOXIA LIU

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2010

SEMI-EMPIRICAL LIKELIHOOD CONFIDENCE INTERVALS FOR THE ROC CURVE WITH MISSING DATA

by

XIAOXIA LIU

Committee Chair:	Dr. Yichuan Zhao
Committee:	Dr. Yixin Fang
	Dr. Jiawei Liu
	Dr. Xu Zhang

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
August 2010

for my family, friends and mentors

ACKNOWLEDGEMENTS

First and foremost I would like to thank my advisor, Dr. Yichuan Zhao, for all his guidance and supports throughout my graduate studies in Georgia State University. I appreciate the opportunity he gave me as a research assistant and finding my thesis topic. He has supported me throughout the thesis with patience and knowledge whilst allowing me the room and freedom to work in my own way.

I would also like to thank my committee members Dr. Xu Zhang, Dr. Jiawei Liu, and Dr. Yixin Fang for accepting my invitation and taking their time to read my thesis and giving me valuable suggestions. Besides, I would like to acknowledge all the professors in the Mathematics and Statistics Department at Georgia State University who taught me classes. They all assisted and inspired me to develop my statistical knowledge and skills.

Moreover, I would like to utilize this special chance to express my deepest gratitude to my parents and brother for supporting me throughout my life and all studies. Thanks for the constant and unconditional love which has been my greatest strength.

Last but not least, the greatest thanks go to my loving, supportive, encouraging husband Jing Qian. This thesis would not have been possible without your patience, help and support. Your constant inspiration and guidance kept me focused and motivated. You always take my problem as your own and encourage me to achieve a higher level. It is so wonderful to have you beside me, in the past, present and future.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 ROC Curve

In medical researches, the receiver operating characteristic (ROC) curve analysis has been extensively used in the evaluation of diagnostic tests and is currently the best-developed statistical tool for describing the performance of such test. ROC curves provide a comprehensive and visually attractive way to summarize the accuracy of predictions. Generally speaking, ROC curve is a graphical plot of the *sensitivity*, or true positives, versus $(1 - \textit{specificity})$, or false positives. It has been in use for years, which was first developed during World War II for signal detection. Its potential for medical diagnostic testing was recognized as early as 1960, although it was in the early 1980s that it became popular, especially in radiology (Pepe, 2003). Nowadays, ROC curves enjoy broader applications in medicine (Shapiro, 1999).

Define a binary test from the continuous test result T as positive if $T \geq c$, negative if $T < c$ using a threshold c. Let D denote the disease status with

$$D = \begin{cases} 1, & \text{if diseased,} \\ 0, & \text{if non-diseased.} \end{cases}$$

Define the corresponding true and false positive fractions at c to be TPF(c) and

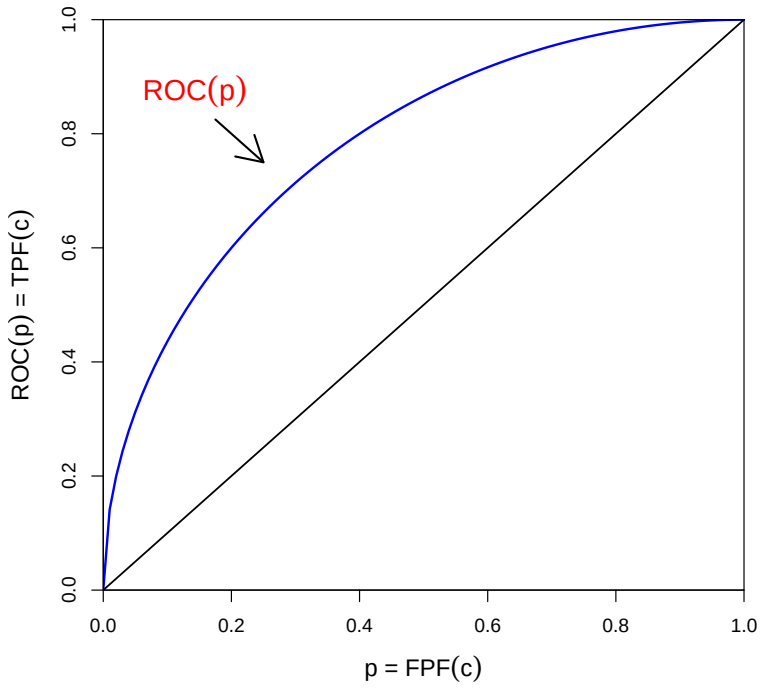

Figure 1.1. An example of ROC curve.

FPF(c), respectively, where TPF(c) = Pr($T \geq c \mid D = 1$), FPF(c) = Pr($T \geq c \mid D = 0$). The ROC curve is the entire set of possible true and false positive fractions attained by dichotomizing T with different thresholds (Pepe, 2003). That is, the ROC curve is

$$\text{ROC}(\cdot) = \{\, (\text{FPF}(c), \text{TPF}(c)), \quad c \in (-\infty, \infty) \,\}.$$

As seen, when c increases, both FPF(c) and TPF(c) decrease. For extreme case, when $c = \infty$, we can get $\lim_{c\to\infty}\text{TPF}(c) = 0$ and $\lim_{c\to\infty}\text{FPF}(c) = 0$. On the other hand, if $c = -\infty$, we have $\lim_{c\to-\infty}\text{TPF}(c) = 1$ and $\lim_{c\to-\infty}\text{FPF}(c) = 1$. Thus, the ROC curve is actually a monotone increasing function in the positive quadrant. See Figure 1.1 for an illustration.

Now, we discuss an alternative way to represent ROC curve. When considering

the results of a particular test in two populations, one population with disease, the other population without disease (well population), we will rarely observe a perfect separation between the two groups. In fact, the distributions of the test results will overlap, as shown later in Figures 4.1 and 4.2. For every possible cut-off value or criterion value selected to discriminate between the two populations, there will be some cases with the disease correctly classified as positive (TPF = True Positive fraction), but some cases with the disease will be classified negative (FNF = False Negative fraction). It is known that if we decrease the false positives, the true positives also decrease. If the threshold is very high, then there will be almost no false positives, but we will not really identify many true positives either. For a continuous-scale diagnostic test, let X be the test results from diseased subjects, and let Y be the test results from non-diseased subjects. At a given cutoff point or threshold c, the sensitivity and specificity are defined as $\text{Se} = \Pr(X \geq c)$ and $\text{Sp} = \Pr(Y < c)$ respectively. If $F(\cdot)$ is the distribution function of X and $G(\cdot)$ is the distribution function of Y, the sensitivity and specificity can then be written as $\text{Se} = 1 - F(c)$ and $\text{Sp} = G(c)$. Then the ROC curve is actually a plot of $1 - F(c)$ versus $1 - G(c)$, for $-\infty < c < \infty$. At a fixed level $q = (1 - \text{specificity})$, the ROC curve can be represented by

$$\Delta_q = 1 - F\{G^{-1}(1 - q)\}, \quad \text{for } 0 < q < 1, \tag{1.1}$$

where G^{-1} is the inverse function of G, i.e., $G^{-1}(q) = \inf\{c : G(c) \geq q\}$.

The ROC curves have been studied for decades. Varieties of approaches regarding estimation of ROC curve have been developed, both parametric and non-parametric. Tosteson and Begg (1988) as well as Goddard and Hinberg (1990) propose ways to model F and G parametrically. Zweig and Campbell (1993) later on provide an extensive review of parametric methods for the ROC curve. The parametric methods

use standard statistical approaches such as maximum likelihood to make inference. To avoid the misspecification problem, non-parametric methods have also been developed to estimate the ROC curve. Refer to Gastwirth and Wang (1988), Hollander and Korwar (1982) and Li, Tiwari and Wells (1996) for examples of comparing two unknown continuous distributions based on independent samples. Especially, for small or moderate sample sizes, the normal approximation may not be applicable because the covariance of the proposed estimator is hard to get or complicated to implement (Liang and Zhou, 2008). To avoid these deficiencies, the empirical likelihood (EL) based method can be used for inference of Δ_q as introduced in the above paragraph.

1.2 Empirical Likelihood

Empirical Likelihood is a nonparametric way of inference based on a data-driven likelihood ratio function. The inference made by EL method does not require the data come from a known family of distributions. Empirical likelihood can be thought of as a bootstrap that does not resample, and as a likelihood without parametric assumptions (Owen, 2001). The original idea of empirical likelihood can date back to Hartley and Rao (1968) in sample survey context and to the nonparametric likelihood ratio inferences for the survival function as described in Thomas and Grunkemeier (1975). Empirical likelihood has been developed by many researchers, and is still undergoing active development. Owen (1988, 1990, 2001) has made systematic studies of the empirical likelihood approach in complete data settings. Hall and La Scala (1990) and DiCiccio *et al.* (1991) develop the empirical likelihood regions. Qin (1994, 1999) also contribute systematically to the empirical likelihood ratio principle. Qin and Lawless (1994) propose an empirical likelihood for parameter defined by general estimating equations and established the Wilks theorem for the empirical likelihood

ratio.

The empirical likelihood approach has many advantages over competitors. The most appealing features include improvement of the confidence region, an increase of accuracy of coverage because of using auxiliary information (Owen, 2001) and easy implementation. The empirical likelihood combines the reliability of the nonparametric methods with the flexibility and effectiveness of the likelihood approach.

We first outline empirical likelihood and related theorem as discussed by Owen (1988, 1990, 2001).

For a random variable $X \in \mathbb{R}$, the cumulative distribution function (CDF) is defined as the function $F(x) = \Pr(X \leqslant x)$, for $-\infty < x < \infty$. We use $F(x-)$ to represent $Pr(X < x)$, so that $Pr(X = x) = F(x) - F(x-)$. Let $I(\cdot)$ be the indicator function, *the empirical cumulative distribution function (ECDF)* of $X_1, ..., X_n$ is

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x), \quad \text{for} \quad -\infty < x < \infty.$$

It is known that the parametric likelihood function for a set of n independent observations $X_1, ..., X_n \sim f(x)$ is

$$\ell(X) = \prod_{i=1}^{n} f(X_i).$$

Refer to the theorem in Owen (2001), let $X_1, ..., X_n \in \mathbb{R}$ be *independent random variables* with a common CDF F_0, and let F_n be their ECDF and F be any CDF, $\ell(F)$ is the nonparametric likelihood of the CDF F. If $F \neq F_n$, then $\ell(F) < \ell(F_n)$, which means the ECDF is the nonparametric maximum likelihood estimate (NPMLE) of F. The empirical likelihood function of the CDF F is

$$\ell(F) = \prod_{i=1}^{n} (F(X_i) - F(X_i-)) = \prod P_i$$

In parametric inference, Wilks's theorem proves that $-2\log(L(\eta_0)/L(\hat{\eta}))$ tends to a chi-squared distribution as $n \to \infty$, which allows us to decide how small $L(\eta)$ must be in order to for η to get rejected. The degree of freedom in the chi-squared distribution usually takes the value of the dimension of the set of η. If we want to get a confidence region for θ we take the image of a confidence region for η, which is

$$\{\theta(\eta)|L(\eta) \geqslant cL(\widehat{\eta})\},$$

where the threshold c is chosen according to Wilks's theorem, with degree of freedom equals the dimension of the set of θ values.

Similarly, we may also use ratios of the nonparametric likelihood for hypothesis test and confidence intervals. For a distribution F, we can define a likelihood ratio

$$R(F) = \frac{\ell(F)}{\ell(F_n)},$$

through the nonparametric likelihood $\ell(F)$ defined above. When there are no ties in the data, the likelihood ratio is

$$R(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^{n} np_i.$$

When there are some ties, the likelihood ratio is

$$R(F) = \prod_{j=1}^{k} \left(\frac{p_j}{\hat{p}_j}\right)^{n_j} = \prod_{j=1}^{k} \left(\frac{np_j}{n_j}\right)^{n_j},$$

where k is the number of distinct values in the data set. Suppose we are interested

in a parameter $\theta = T(F)$ for some function T of distributions. Define the profile likelihood ratio function as

$$\mathcal{R}(\theta) = \sup\{R(F) \mid T(F) = \theta, F \in \mathcal{F}\}.$$

For some threshold value r_0, the empirical likelihood confidence regions are of the form

$$\{\theta \mid \mathcal{R}(\theta) \geq r_0\}.$$

Owen (1988) shows the analogue of Wilks theorem for convergence of the empirical likelihood ratio for the population mean. Let $X_1, ..., X_n$ be independent random variables with non-degenerate distribution function F_0 with $\int |x|^3 dF_0 < \infty$. For positive $c < 1$, let

$$\mathcal{F}_{c,n} = \{F \mid R(F) \geq c, F \ll F_n\},$$

and define $X_{U,n} = \sup \int x dF$ and $X_{L,n} = \inf \int x dF$ with both extreme take over $F \in \mathcal{F}_{c,n}$. Then as $n \to \infty$,

$$\Pr\{X_{L,n} \leqslant E(X) \leqslant X_{U,n}\} \to \Pr(\chi^2_{(1)} \leqslant -2 \log c).$$

Owen (1990,2001) proves a remarkable result similar to Wilks theorem,which is known as the Empirical Likelihood Theorem(ELT). Let $X_1, ..., X_n$ be independent random variables with common distribution F_0, which has mean $\mu_0 = E(X_i)$ and variance $0 < Var(X_i) < \infty$.

$$-2 \log(\mathcal{R}(\mu_0)) \to \chi^2_{(1)}$$

as $n \to \infty$, where the profile empirical likelihood ratio function for the mean is

$$\mathcal{R}(\mu_0) = \max \left\{ \prod_{i=1}^{n} nw_i | \sum_{i=1}^{n} w_i x_i = \mu_0, w_i \geqslant 0, \sum_{i=1}^{n} w_i = 1 \right\}.$$

The resulting empirical likelihood confidence region of the mean with the form

$$\{\mu | \mathcal{R}(\mu) \geqslant r_0\} = \left\{ \sum_{i=1}^{n} w_i X_i | \prod_{i=1}^{n} nw_i \geqslant r_0, w_i \geqslant 0, \sum_{i=1}^{n} w_i = 1 \right\}.$$

The fact that empirical likelihood ratio has a limiting chi-squared distribution, leading to tests and confidence intervals for a variety of statistical problems. For example, we may notice applications in linear models (Owen, 2001; Chen, 1993, 1994), generalized linear models (Kolaczyk 1994) and general estimating equation (Qin and Lawless (1994)). Chen and Hall (1993) introduced smoothed empirical likelihood based confidence intervals for quantiles on one population. In the next chapter, we are going to review the literatures using empirical likelihood ratio statistic to make inference for ROC curves.

1.3 Missing Data and Hot Deck Imputation

When making statistical inference, it is usually assumed that all responses in the sample are available. However, this may be violated in many practical situations. The responses may be missing for various reasons, such as subject's refusal to answer an item, loss of information caused by uncontrollable factors, failure to collect correct information and so on. In fact, missing data are very common in opinion polls, marketing surveys, questionnaires, socioeconomic investigations, medical researches and other scientific studies (Wang and Rao (2002)). In statistics, missing values occur when no data value is stored for the variable in the current observation. Since

missing values can badly distort the findings of research, when it occurs, the usual inferential procedures for complete data sets cannot be applied directly. It is critical to handel missing data properly. There are several useful distinctions we can make for the types of missing data. If the data are Missing Completely At Random (MCAR), then missing values cannot be predicted any better with the information in the data matrix, observed or not. In another words, MCAR happens when the probability that an observation X_i is missing is unrelated to the value of X_i or to the value of any other variables. Often, data are not missing completely at random, they may be classified as Missing At Random (MAR), if the probability that a cell is missing may depend on the observed data, but after controlling for observed data, that probability must be independent of unobserved data. For example, a group of people were asked for their vote choice, but there are some missing values in this variable. At the same time, there is another variable gender that was recorded. The process of missing in vote choice is missing completely at random if, say, an individual's decision whether to answer the question is based on flipping a coin. On the other hand, the process of missing in vote choice is missing at random if, say, female people are more likely to refuse to answer the vote choice question than the male. If data are not missing at random or completely at random then they are classed as Missing Not at Random (MNAR).

In statistical analysis, we may define the complete data as Y=(y_{ij}) and the missing-data indicator matrix M=(M_{ij}). The underlying missing data mechanism is characterized by the conditional distribution of M given Y, which is equivalent to $f(M \mid Y, \phi)$, where ϕ denotes unknown parameters. If the data is MCAR, we have

$$f(M \mid Y, \phi) = f(M \mid \phi) \quad \text{for all } Y, \phi.$$

Let Y_{obs} denote the observed values of Y, and Y_{mis} be the missing components. Then MAR can be expressed by:

$$f(M \mid Y, \phi) = f(M \mid Y_{obs}, \phi) \quad \text{for all } Y_{mis}, \phi.$$

There are a number of ways or treatments dealing with missing data. Little and Rubin (2002) summarized and grouped the current available methods into four categories, which are not mutually exclusive. The main methods proposed are procedures based on completely recorded units, weighting procedure, imputation-based procedure, and model-based procedure. For the imputation-based procedure, the missing values are filled in and the resultant complete data are then analyzed using standard statistical methods as if they were true observations. One commonly used imputation procedures is *hot deck* imputation, where recorded units in the sample are used to substitute values. The hot deck literally refers to the deck of matching computer cards for the donors available for a nonrespondents (Little and Rubin, 2002). It goes back over 50 years and was used quite successfully by the Census Bureau, survey and others. With this specific imputation method, we replace missing values by values from similar responding units in the sample.

Current available methods in analyzing ROC curves are limited to complete data regardless of parametric or non-parametric settings. The empirical likelihood method needs modifications when dealing with data with missing or imputed values. We want to extend the previous study and concern the situation that one model is parametric and another one is non-parametric, both with missing data in them. This leads to a semi-parametric two-sample model. In this thesis, we are interested in constructing the confidence intervals for the ROC curves, or Δ_q with missing data under this specific context by using empirical likelihood ratio methods.

1.4 Structure

The structure of the remainder of thesis is as follows. Chapter 2 presents the literature review of empirical and semi-empirical likelihood methods with applications to ROC curves. Chapter 3 shows inference procedure. We introduce the hot deck imputation method first. Then the smoothed empirical likelihood is generated for the ROC curve. Also, the semi-empirical likelihood based confidence interval is constructed and the asymptotic results are established. The semi-empirical likelihood ratio statistic and the corresponding likelihood equations are developed in this chapter. In Chapter 4, simulation studies are conducted to evaluate the finite sample performance of the proposed method. Finally, Chapter 5 gives a summary and discussion as well as the description for future studies. All technical details and proofs are included in the Appendix A. R code for the simulation studies is attached in Appendix B.

Chapter 2

LITERATURE REVIEW

2.1 Empirical Likelihood Ratio Confidence Interval for the ROC Curve

The empirical likelihood principle has been widely used in statistical literature. In this chapter, we look back to the previous work using empirical likelihood smoothing strategy to obtain a confidence interval for the ROC curve Δ_q. Development of confidence intervals of an ROC curve has received much attention because it is more important than point estimates and more useful for practitioners in making diagnostic decisions (Su, Qin and Liang, 2009). The first use of empirical likelihood ratio function to get confidence intervals appears to be Thomas and Grunkemeier (1975). They show that empirical likelihood ratio confidence intervals for a survival probability based on the $\chi^2_{(1)}$ distribution have asymptotically correct coverage levels. Later on, the empirical likelihood methods for constructing confidence regions for the mean parameter of the population were developed systematically by Owen (1988) and Owen (1990). In general, we can see it as a nonparametric or semi-parametric version of Wilks' Theorem and a multivariate generalization of the work by Owen (1988). There are two advantages of this empirical likelihood formulation. One is that the information contained in the zero observations is fully utilized. The other is that the proposed confidence intervals are more reflective to the likely situation that the non-zero value distribution is skewed (Chen and Qin, 2003). Comparing to

normal approximation method and bootstrap method, EL method can improve the confidence region, and increase the accuracy of the coverage (Hall and La Scala, 1990). Hsieh and Turnbull (1996) estimate the ROC curve empirically or non-parametrically. They show that the empirical ROC curve estimator shares the same good asymptotic properties with standard maximum likelihood estimators. To further improve the coverage of the empirical likelihood confidence interval, Chen and Qin (2003) proposed an empirical Bartlett correction to the empirical likelihood confidence intervals based on the bootstrap. Recently, much attention has been paid to semi-parametric inference. Li, Tiwari and Wells (1999) show a semi-parametric way to estimate the ROC curves. They study two sample inference through the quantile comparison function $G\{F^{-1}(p)\}$ assuming that G is known and F is unknown. Zhou, McClish and Obuchowski (2002) give a comprehensive survey of the estimation methods for ROC curves which gives us a summary of the currently popular methods. Also, smoothing strategies or techniques are developed to correct the discontinuity of the ROC curves. Chen and Hall (1993) first of all introduced smoothed empirical likelihood-based confidence intervals for quantiles on one population. Zou, Hall and Shapiro (1997), Lloyd (1998) and Ren, Zhou and Liang (2004) proposed various smoothed estimators for ROC curves among others. They construct a smooth estimator of $R(t)$ by considering $\tilde{R}(t) = 1 - \tilde{F}_{1n_1}\{\tilde{F}^{-1}_{2n_2}(1-t)\}$, where the $\tilde{F}_{in_i}$ are smooth versions of F_{in_i}, for example, kernel distribution estimators. Zhou and Jing (2003) developed smoothed empirical likelihood confidence intervals for the difference of quantiles. Claeskens *et al.* (2003) proposed a smoothed empirical likelihood method for confidence intervals of ROC curves. Two concerns with their approach are that bandwidths have to be selected and the computation of their approach may be very expensive which may not be feasible in practice. The principle of our smoothed empirical likelihood is similar to that of Chen and Hall (1993) and Claeskens *et al.* (2003) in spirit.

Current available methods in analyzing ROC curves are limited to complete data regardless of parametric or non-parametric settings. The empirical likelihood method needs modifications when dealing with data with missing or imputed values. Wang and Rao (2002) employ the empirical likelihood method to construct confidence intervals for the mean of the dependent variable in a nonparametric regression model with missing data. Zhou and Liang (2005) extend the study to semi-parametric inference for ROC curves with censoring. To remedy the deficiency of discontinuity of the ROC curves, Liang and Zhou (2008) further propose a smoothed semiparametric likelihood-based confidence intervals approach for ROC curves when the observations are censored. They present combining smoothing technique and the approaches developed by Zhou and Liang (2005) to derive an appropriate estimating equation because the naive estimating function is non-differentiable and the corresponding estimating equation is inconsistent. Qin and Zhang (2009) investigate the semi-empirical likelihood confidence intervals for the quantiles differences of two population with missing data.

2.2 Contributions

The procedure in our context is different from that for usual situations. In this thesis, we want to extend the previous studies and concern the situation that when one model is parametric while the other one is non-parametric, both with missing data in them. As a matter of fact, this is a very common case in medical research or related fields. Say, when comparing a new treatment with control treatment, we tend to have more if not enough information about the well developed treatment (i.e. control treatment), while the new treatment is less known. This leads to a semi-parametric two-sample model, which can reflect the difference of two samples of missing data. Let

X and Y be the responses of two samples, for example, the diseased and non-diseased subjects, and we assume $F(.)$ and $G(.)$ are the distribution functions of X and Y respectively. Furthermore, we have the assumption that the population distribution function F is non-parametric while G is parametric, and both X and Y with missing data in them. In this thesis, we are interested in constructing the confidence intervals for the ROC curves, or Δ_q with missing data under this specific context by using empirical likelihood ratio methods to avoid estimating the covariance matrix and using normal approximation (Su *et al.*, 2009). We are interested in establishing asymptotic distribution of the resulting statistics and deriving the empirical likelihood-based confidence intervals of the parameters of interest under mild assumptions. We also prove that the resulting log likelihood ratio is still asymptotically scaled chi-square distributed under such conditions.

Chapter 3

INFERENCE PROCEDURE

3.1 Missing Data Imputation

Consider the following simple random samples of incomplete data associated with populations (x, δ_x) and (y, δ_y):

$$(x_i, \delta_{xi}), i = 1, ..., m; \quad (y_j, \delta_{yj}), j = 1, ..., n,$$

where missing data indictor

$$\delta_{xi} = \begin{cases} 0, & \text{if } x_i \text{ is missing,} \\ 1, & \text{otherwise.} \end{cases}$$

$$\delta_{yj} = \begin{cases} 0, & \text{if } y_j \text{ is missing,} \\ 1, & \text{otherwise.} \end{cases}$$

Throughout this thesis we assume MCAR, i.e. $P(\delta_x = 1|x) = P_1$(constant) and $P(\delta_y = 1|y) = P_2$ (constant). Furthermore, we have the assumption that (x,δ_x) and $(y,\ \delta_y)$ are independent. Let $r_x = \sum_{i=1}^{m} \delta_{xi}$, $r_y = \sum_{j=1}^{n} \delta_{yj}$, $m_x = m - r_x$ and $m_y = n - r_y$. The respondents with respect to x and y can be written as s_{rx} and s_{ry}, respectively, while the non-respondents are denoted as s_{mx} and s_{my} corresponding to

x and y respectively. The means of the respondents units with respect to x and y are expressed as

$$\bar{x}_r = \frac{1}{r_x} \sum_{i \in s_{rx}} x_i,$$

$$\bar{y}_r = \frac{1}{r_y} \sum_{i \in s_{ry}} y_i.$$

In this thesis, we choose random hot deck imputation method to impute the missing values. Let x_i^* and y_j^* denote the imputed value for the missing data with respect to x and y, respectively. For the sample X, which comes from a non-parametric population, we impute the missing values by selecting simple random samples from the observed ones. We select a simple random sample of size m_x with replacement from s_{rx} and then use the associated x values as donors, that is, $x_i^* = x_j$ for some $j \in s_{rx}$. For the sample Y, which comes from a parametric population, we first get Maximum Likelihood Estimator (MLE) of population parameter, then select simple random samples from the population with this estimated population parameter. Let $\hat{\theta}$ denote MLE of θ from the sample $\{y_j, j \in s_{ry}\}$. Then we select a simple random sample of size m_y with replacement from the population $G_{\hat{\theta}}(\cdot)$. Following this procedure, we obtain y_j^*. As shown, the 'complete' data set after imputation is:

$$x_{I,i} = \delta_{xi} x_i + (1 - \delta_{xi}) x_i^*, \quad y_{I,j} = \delta_{yj} y_j + (1 - \delta_{yj}) y_j^*,$$

where $i = 1, ..., m, \quad j = 1, ..., n.$

3.2 Smoothed Semi-empirical Likelihood

In this section, we develop a semi-empirical likelihood-based confidence interval for the ROC Curve Δ_q by using kernel smoothing technique (Chen and Hall, 1993),

where $\Delta_q = 1 - F\{G^{-1}(1-q)\}$ $(0 < q < 1)$ as defined in equation (1.1).

Given the samples $\{(x_{I,i}, \delta_{xi}), i = 1, ..., m\}$ and $\{(y_{I,j}, \delta_{yj}), j = 1, ..., n\}$, the corresponding semi-likelihood function can be written as

$$\prod_{i=1}^{m} p_i \prod_{j=1}^{n} g_\theta(y_{I,j}) = \prod_{i=1}^{m} p_i \prod_{j=1}^{n} \{g_\theta^{\delta_{yj}}(y_j) g_{\hat{\theta}}^{1-\delta_{yj}}(y_{I,j})\}, \tag{3.1}$$

where

$$\sum_{i=1}^{m} p_i = 1, \quad p_i > 0, \;\; i = 1, ..., m \;\text{ and }\; \theta \in \Theta. \tag{3.2}$$

To introduce the smoothed semi-empirical likelihood ratio statistic, we first define $H(t) = \int_{-\infty}^{t/h} K(u)du$, where $K(u)$ is kernel function satisfying some conditions stated later in asymptotic studies, and $h = h_n > 0$ is a sequence of bandwidths with $h_n \to 0$ and $nh_n \to 0$ as $n \to \infty$. We write

$$\omega(x_{I,i}, \theta, \Delta_q) = H\{G_\theta{}^{-1}(1-q) - x_{I,i}\} - (1 - \Delta_q).$$

Since the function (3.1) attains its maximum value over $\{p_i, \theta\}$ satisfying (3.2) when $p_i = m^{-1}\,(i = 1, ..., m)$ and $\theta = \hat{\theta}$, it follows from Chen and Hall (1993), Qin (1994) and Qin (1997) that the semi-empirical likelihood ratio statistic can be defined as

$$\begin{aligned} \mathcal{R}(\Delta_q, \theta) &= \sup_{p_1,...,p_m,\theta} \frac{\prod_{i=1}^{m} p_i \prod_{j=1}^{n} g_\theta(y_{I,j})}{m^{-m} \prod_{i=1}^{m} g_{\hat{\theta}}(y_{I,j})} \\ &= \sup_{p_1,...,p_m,\theta} \prod_{i=1}^{m} mp_i \prod_{j=1}^{n} g_\theta^{\delta_{yj}}(y_j) \left[\prod_{i=1}^{m} g_{\hat{\theta}}^{\delta_{yj}}(y_j)\right]^{-1}, \end{aligned}$$

where $p_1, ..., p_m$ are subject to restrictions

$$\sum_{i=1}^{m} p_i = 1, \quad \sum_{i=1}^{m} p_i \omega(x_{I,i}, \theta, \Delta_q) = 0, \quad p_i > 0, \;\; i = 1, ..., m \tag{3.3}$$

Following Qin (1994) and Qin (1997), we write

$$\mathcal{R}(\Delta_q, \theta) = \sup_{\theta} \left\{ \sup_{p_1,...,p_m} \prod_{i=1}^{m} m p_i \prod_{j=1}^{n} g_{\theta}^{\delta_{yj}}(y_j) \left[\prod_{i=1}^{m} g_{\hat{\theta}}^{\delta_{yj}}(y_j) \right]^{-1} \right\}$$

and consider the following maximization problem at fixed θ,

$$H(\Delta_q, \theta) = \max_{p_1,...,p_m} \left\{ \sum_{i=1}^{m} \log m p_i + \sum_{j=1}^{n} \log g_{\theta}^{\delta_{y_j}}(y_j) \;\middle|\; \text{restrictions (3.3)} \right\}. \tag{3.4}$$

By the method of Langrange multipliers, the maximization problem of (3.4) can be formulated as

$$H = \sum_{i=1}^{m} \log m p_i + \sum_{j=1}^{n} \log g_{\theta}^{\delta_{y_j}}(y_j) + \gamma \left(1 - \sum_{i=1}^{m} p_i \right) - m\lambda(\theta) \sum_{i=1}^{m} p_i \omega(x_{I,i}, \theta, \Delta_q).$$

Then,

$$\begin{aligned}
&\frac{\partial H}{\partial p_i} = p_i^{-1} - \gamma - m\lambda(\theta)\omega(x_{I,i}, \theta, \Delta_q) = 0 \\
&\qquad \Rightarrow p_i = \frac{1}{\gamma + m\lambda(\theta)\omega(x_{I,i}, \theta, \Delta_q)}; \\
&\sum_{i=1}^{m} \left(p_i \frac{\partial H}{\partial p_i} \right) = m - \gamma \sum_{i=1}^{m} p_i - m\lambda(\theta) \sum_{i=1}^{m} p_i \omega(x_{I,i}, \theta, \Delta_q) = 0 \\
&\qquad \Rightarrow \gamma = m.
\end{aligned}$$

Thus,

$$p_i = \frac{1}{m\{1 + \lambda(\theta)\omega(x_{I,i}, \theta, \Delta_q)\}},$$

and $\lambda(\theta)$ is determined by the following equation,

$$\frac{1}{m}\sum_{i=1}^{m} \frac{\omega(x_{I,i}, \theta, \Delta_q)}{1 + \lambda(\theta)\,\omega(x_{I,i}, \theta, \Delta_q)} = 0, \tag{3.5}$$

and

$$H(\Delta_q, \theta) = -\sum_{i=1}^{m} \log\{1 + \lambda(\theta)\omega(x_{I,i}, \theta, \Delta_q)\} + \sum_{j=1}^{n} \log g_\theta^{\delta_{y_j}}(y_j).$$

Let $\partial H(\Delta_q, \theta)/\partial\theta = 0$, we can obtain the semi-empirical likelihood equation as follows,

$$\lambda(\theta) \sum_{i=1}^{m} \frac{h^{-1}K[\{G_\theta^{-1}(1-q) - x_{I,i}\}/h]}{1 + \lambda(\theta)\omega(x_{I,i}, \theta, \Delta)}\,\alpha(\theta) = \sum_{j=1}^{n} \delta_{yj} \frac{\partial \log g_\theta(y_j)}{\partial\theta}, \tag{3.6}$$

where

$$\alpha(\theta) = -\frac{1}{g_\theta({G_\theta}^{-1}(1-q))} \int_{-\infty}^{{G_\theta}^{-1}(1-q)} \frac{\partial g_\theta(t)}{\partial\theta}\,dt.$$

3.3 Asymptotic Studies

3.3.1 Assumptions

Let θ_0 denote the true value of θ. We make the following assumptions (i) to (vi) on the distribution of $G_\theta(y)$. Meanwhile, we make some additional assumptions.

(i) $\theta_0 \in \Theta$ and Θ is an open interval.

(ii) The distribution of $G_\theta(y)$ has a common support so that the set $A = y : g_\theta(y) > 0$ is independent of θ.

(iii) For every $y \in A$, the density function $g_\theta(y)$ is three times differentiable with respect to θ.

(iv) The integral $\int g_\theta(y)dy$ can be differentiated twice under the integral sign. For any $\theta \in \Theta, \quad g_\theta(G_\theta^{-1}(q)) \neq 0$, $\alpha''(\theta)$ exists and is continuous in a neighborhood of θ_0, and $\alpha(\theta_0) \neq 0$.

(v) The Fisher information matrix $I(\theta)$ with the entries $I(\theta) = E_\theta\{\partial \log g_\theta(y)/\partial\theta\}^2 = -E_\theta\{\partial^2 \log g_\theta(y)/\partial\theta^2\}$ which is positive definite, $0 < I(\theta) < \infty$.

(vi) $|(\partial^3/\partial\theta^3)\log g_\theta(y)| < M(y)$, for all $y \in A$, $\theta_0 - c < \theta < \theta_0 + c$ (for some c), with $E_{\theta_0}\{M(y)\} < \infty$.

(vii) There exists a constant $t_0 \geq 2$ such that $f^{t_0-1}(\cdot)$ exists and is continuous in a neighborhood of $F^{-1}(1-\Delta_q)$ with $f\{F^{-1}(1-\Delta_q)\} > 0$, where f is the probability density function of X.

(viii) $\frac{n}{m} \to k(0 < k < \infty)$ as $m, n \to \infty$.

(ix) The kernel function $K(u)$ is bounded and satisfies Lipschitz condition of order 1; $K^{(2)}(u)$ exists and is bounded. Assume that for some $C > 0$,

$$\int_{|u|>C/h^{t_0}} K(u)du = O(h^{t_0}), \quad \int |u^{t_0}K(u)|du < \infty,$$

and that $K(u)$ has finite support satisfying

$$\int u^j K(u)\,du = \begin{cases} 1, & \text{if } j = 0, \\ 0, & \text{if } 1 \leq j \leq t_0 - 1. \end{cases}$$

(x) There exists an r $(1/3 < r < 1/2)$ such that $n^r h^{t_0} \to 0$ and $n^r h \to \infty$.

(xi) $\sqrt{n}(\hat{\theta} - \theta_0) = O_p(1)$ and $\partial l_2(\hat{\theta})/\partial\theta = 0$, where $l_2(\theta) = (1/n)\sum_{j=1}^n \delta_{yj} \log g_\theta(y_j)$.

3.3.2 A Wilks' Theorem

Theorem 3.1. *Suppose that assumptions (i) to (xi) are satisfied, then with probability tending to 1, there exists a root* $\hat{\theta_{EL}}$ *of equation (3.6) such that* $\mathcal{R}(\Delta_q, \theta)$ *attains its local minimum at* $\hat{\theta}_{EL}$ *and as* $m, n \to \infty$,

$$-2\rho(\Delta_q, \theta_0) \log \mathcal{R}(\Delta_q, \hat{\theta}_{EL}) \xrightarrow{d} \chi^2_{(1)},$$

where

$$\rho(\Delta_q, \theta) = \frac{kP_2 I(\theta)\Delta_q(1-\Delta_q) + \beta_0^2(\Delta_q, \theta)}{kP_2 I(\theta)(1 - P_1 + P_1^{-1})\Delta_q(1-\Delta_q) + \beta_0^2(\Delta_q, \theta)},$$

and

$$\beta_0 = \alpha(\theta_0) f(F^{-1}(1-\Delta_q)),$$

Theorem 3.1 implies that the asymptotic distribution of the log-semi-empirical likelihood ratio statistic is scaled chi-square variable. According to Qin and Zhang (2009), the reason for this deviation away from the standard chi-square is because the complete data after imputation are dependent instead of independent.

To construct a confidence interval on Δ_q using Theorem 3.1, we need to get a consistent estimator of $\rho(\Delta_q, \theta_0)$. The response rates P_1 and P_2 can be consistently estimated by $\hat{P}_1 = \frac{1}{m}\sum_{i=1}^{m} \delta_{xi}$ and $\hat{P}_2 = \frac{1}{n}\sum_{j=1}^{n} \delta_{yj}$ respectively. Also, k is estimated by n/m. Similar to the proof of Lemma A.2 in Appendix A, we can obtain an estimator for $\beta_0(\Delta_q, \theta_0)$,

$$\hat{\beta}(\Delta_q, \theta_0) = \frac{1}{mh}\sum_{i=1}^{m} K((G^{-1}_{\hat{\theta}_{EL}}(1-q) - x_{I,i})/h)\alpha(\hat{\theta}_{EL}).$$

Moreover, $\hat{I}(\theta_0) = I(\theta_{EL})$ are consistent estimator of $\beta_0(\Delta_q, \theta_0)$ and $I(\theta_0)$, respectively. The resulting $\rho(\Delta_q, \hat{\theta}_{EL})$ is a consistent estimator of $\rho(\Delta_q, \theta_0)$.

Let t_α satisfy $P(\chi_1^2 \leq t_\alpha) = 1 - \alpha$. Following theorem , the semi-empirical likelihood based confidence interval for the ROC curve Δ_q with asymptotically correct coverage probability $1 - \alpha$ can be established as:

$$\{\Delta_q : -2\hat{\rho}(\Delta_q, \hat{\theta}_{EL}) \log \mathcal{R}(\Delta_q, \hat{\theta}_{EL}) \leq t_\alpha\}.$$

Chapter 4

NUMERICAL STUDIES

4.1 Monte Carlo Simulation

In this chapter, we conducted extensive simulation studies to investigate the finite sample performance of the proposed semi-empirical likelihood based confidence intervals for the ROC curve Δ_q, especially with small and moderate sample sizes.

Two setups were considered in the simulation studies. In setup (I), the diseased population X followed a normal distribution with mean 1 and variance 1, while the non-diseased population Y is distributed as the standard normal. Independent random samples x and y are drawn from populations X and Y respectively. We chose four combinations of different sample sizes for x and y, i.e., $(m, n) = (50, 50)$, $(75, 75)$, $(100, 100)$, and $(200, 150)$. Meanwhile, under each combination of sample sizes, we investigated the following response rates for x and y, $(P_1, P_2) = (0.6, 0.7)$, $(0.8, 0.7)$ and $(0.9, 0.8)$. Thus, we were able to perform a comprehensive evaluation of the performance of semi-empirical likelihood confidence intervals. For each scenario of certain missing rates and sample sizes, we generated 1,000 independent random samples of data $\{(x_i, \delta_{xi}), i = 1, ..., m;\ (y_j, \delta_{yj}), j = 1, ..., n\}$, and then constructed the proposed semi-empirical likelihood based confidence intervals on the ROC curve Δ_q at $q = 0.1$ and 0.3 for each sample. The nominal level of the confidence intervals is $1 - \alpha = 0.95$. The setup (II) is the same as the setup (I) except that the

diseased and non-diseased populations are from exponential distribution instead of the normal. We set the densities of X and Y as $f_X(x) = 0.5\exp(-0.5x)\,I(x \geq 0)$, and $f_Y(y) = \exp(y)\,I(y \geq 0)$ respectively, where $I(\cdot)$ is the indicator function. The proposed semi-empirical likelihood based confidence intervals for the ROC curve Δ_q at $q = 0.5$ and 0.7 are constructed for each sample.

As an illustration, Figure 4.1 and Figure 4.2 display the distributions of disease and non-disease populations and the theoretical ROC curves for setups (I) and (II) respectively. The area under the ROC curve in setup (I) equals 0.760, while the area under the ROC curve in setup (II) is 0.667. It can be also seen from the Figures that the area under the ROC curve in setup (I) is larger than that in setup (II), which means it would be easier for a diagnostic test to discriminate those with and without the disease under setup (I) than under setup (II). The points on the ROC curves in the two figures are those that we will construct semi-empirical likelihood based confidence intervals.

In setup (I), we set the Kernel function as $K(u) = (\sqrt{2\pi})^{-1}\exp(-u^2/2)$, and the bandwidth $h = (3/2)\,m^{-1/3}$. The same Kernel function was used in setup (II), but the bandwidth was chosen as $h = m^{-1/3}$.

The simulation study is coded in R software. One major difficulty in the simulation is to solve the estimating equations to obtain the estimated lower and upper bounds of semi-empirical likelihood confidence intervals. This involves solving three estimating equations with three parameters simultaneously (one parameter is the upper or lower bound of confidence interval, the other two are nuisance parameters in the semi-empirical likelihood equations). The R package "BB" (which is developed to solve nonlinear system of equations) is used to do this. An example of R code is attached in Appendix B.

Tables 4.1 and 4.2 display the results of the simulation study under setup (I).

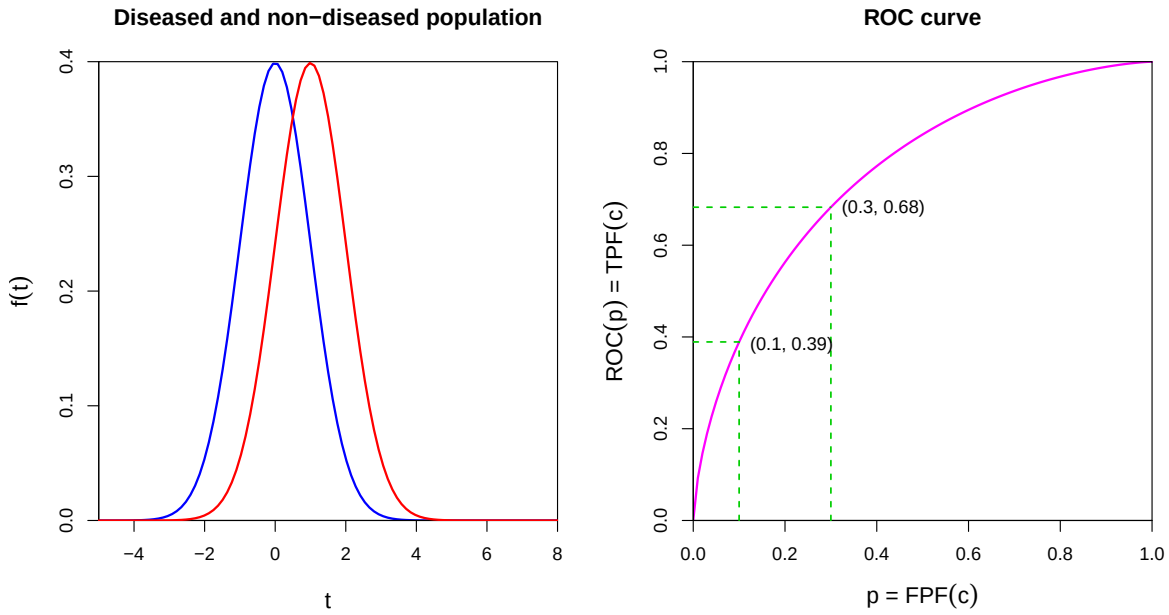

Figure 4.1. Setup (I): disease population (red curve), non-diseased population (blue curve) and theoretical ROC curve (magenta curve).

Each value in the tables is based on the average of 1000 simulations. It can be seen that the coverage probabilities of the semi-empirical likelihood based confidence intervals are very close to the nominal confidence level $1-\alpha = 0.95$ for every response rate and sample size. Meanwhile, the average lengths of the confidence intervals are small. In both tables, we may notice that under the same response rate, the average length (AL) decreases as the sample size increases. The average left endpoint (LE) and right endpoint (RE) are getting closer to the true value of Δ_q as the sample size increases. On the other hand, under the same sample size, the average length (AL) of confidence intervals is getting smaller as the response rate getting larger. The same trend holds for the average left endpoint (LE) and right endpoint (RE) values as they are getting closer to the true Δ_q when response rate increases.

Table 4.3 and Table 4.4 show the results of the simulation corresponding to the

Table 4.1. Semi-empirical likelihood confidence intervals for the ROC curve Δ_q under Setup (I) when $q = 0.1$ $(\Delta_q = 0.3891)$.

(P_1, P_2)	(m, n)	CP(%)	LE	RE	AL
(0.6, 0.7)	(50, 50)	95.6	0.2205	0.6042	0.3837
	(75, 75)	94.7	0.2442	0.5724	0.3281
	(100, 100)	95.3	0.2570	0.5469	0.2899
	(200, 150)	95.3	0.2832	0.5075	0.2242
(0.8, 0.7)	(50, 50)	95.3	0.2350	0.5813	0.3463
	(75, 75)	95.3	0.2531	0.5449	0.2917
	(100, 100)	95.0	0.2701	0.5275	0.2574
	(200, 150)	95.8	0.2938	0.4947	0.2008
(0.9, 0.8)	(50, 50)	95.3	0.2491	0.5694	0.3203
	(75, 75)	94.9	0.2678	0.5366	0.2687
	(100, 100)	94.7	0.2827	0.5205	0.2378
	(200, 150)	94.9	0.3020	0.4876	0.1856

NOTE: : CP(%): coverage probability, LE: the average left endpoint, RE: the average right endpoint and AL: the average length of the interval.

Table 4.2. Semi-empirical likelihood confidence intervals for the ROC curve Δ_q under Setup (I) when $q = 0.3$ $(\Delta_q = 0.6828)$.

(P_1, P_2)	(m, n)	CP(%)	LE	RE	AL
(0.6, 0.7)	(50, 50)	95.6	0.4659	0.8329	0.3670
	(75, 75)	94.5	0.5085	0.8186	0.3100
	(100, 100)	95.0	0.5271	0.8027	0.2756
	(200, 150)	94.9	0.5662	0.7781	0.2119
(0.8, 0.7)	(50, 50)	95.1	0.4896	0.8184	0.3288
	(75, 75)	95.3	0.5200	0.7982	0.2782
	(100, 100)	95.0	0.5417	0.7863	0.2446
	(200, 150)	94.7	0.5737	0.7649	0.1912
(0.9, 0.8)	(50, 50)	94.9	0.4995	0.8035	0.3040
	(75, 75)	94.7	0.5356	0.7906	0.2550
	(100, 100)	95.2	0.5574	0.7825	0.2251
	(200, 150)	94.8	0.5873	0.7628	0.1755

NOTE: : CP(%): coverage probability, LE: the average left endpoint, RE: the average right endpoint and AL: the average length of the interval.

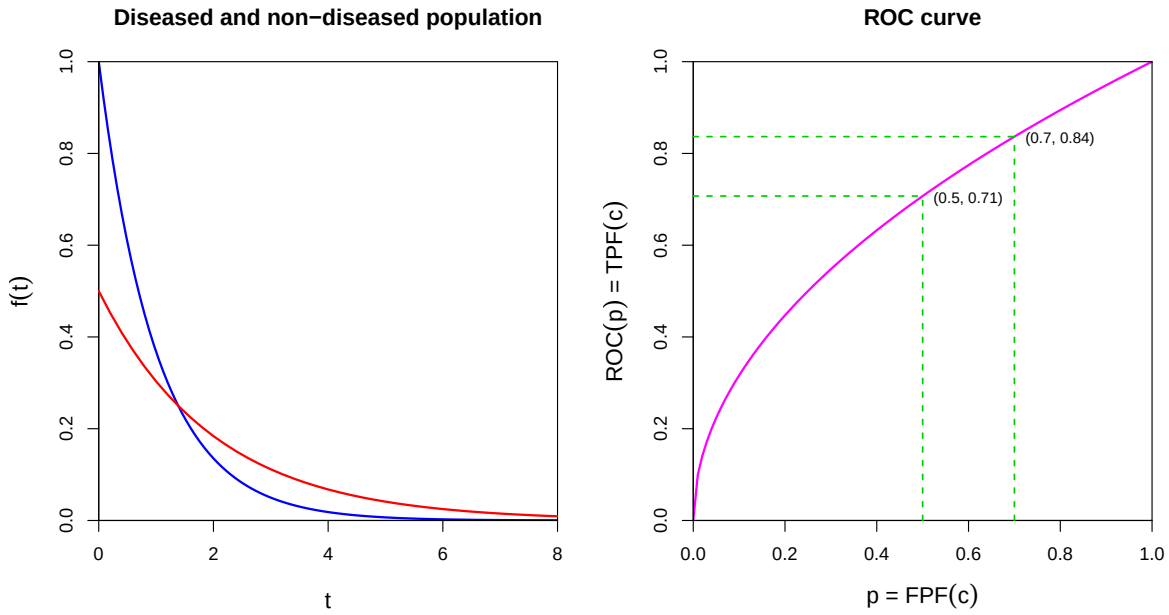

Figure 4.2. Setup (II): disease population (red curve), non-diseased population (blue curve) and theoretical ROC curve (magenta curve).

setup (II). Similar results as those Tables 4.1 and 4.2 can be observed. The coverage probabilities of the confidence intervals based on semi-empirical likelihood are close to the nominal confidence level 0.95 for each combination of the missing rate and sample size. The proposed method works very well even for small sample sizes, such as $(m, n) = (50, 50)$. The average lengths are small and decrease as the sample sizes increase for the same response rate. The consistent result holds for the average lengths when the sample size is the same but the response rate goes up.

Table 4.3. Semi-empirical likelihood confidence intervals for the ROC curve Δ_q under Setup (II) when $q = 0.5 \quad (\Delta_q = 0.7071)$.

(P_1, P_2)	(m, n)	CP(%)	LE	RE	AL
(0.6, 0.7)	(50, 50)	95.4	0.4911	0.8668	0.3757
	(75, 75)	95.7	0.5341	0.8489	0.3148
	(100, 100)	95.1	0.5559	0.8312	0.2753
	(200, 150)	95.6	0.5980	0.8016	0.2036
(0.8, 0.7)	(50, 50)	94.9	0.5205	0.8490	0.3285
	(75, 75)	95.0	0.5547	0.8290	0.2743
	(100, 100)	95.6	0.5791	0.8183	0.2392
	(200, 150)	95.4	0.6146	0.7920	0.1773
(0.9, 0.8)	(50, 50)	95.3	0.5403	0.8427	0.3024
	(75, 75)	95.0	0.5742	0.8246	0.2503
	(100, 100)	94.9	0.5865	0.8068	0.2203
	(200, 150)	95.4	0.6206	0.7841	0.1635

NOTE: : CP(%): coverage probability, LE: the average left endpoint, RE: the average right endpoint and AL: the average length of the interval.

Table 4.4. Semi-empirical likelihood confidence intervals for the ROC curve Δ_q under Setup (II) when $q = 0.7$ $(\Delta_q = 0.8366)$.

(P_1, P_2)	(m, n)	CP(%)	LE	RE	AL
(0.6, 0.7)	(50, 50)	95.7	0.6550	0.9409	0.2859
	(75, 75)	94.8	0.6975	0.9333	0.2359
	(100, 100)	95.7	0.7136	0.9226	0.2090
	(200, 150)	95.6	0.7479	0.9030	0.1550
(0.8, 0.7)	(50, 50)	94.8	0.6839	0.9306	0.2467
	(75, 75)	94.7	0.7150	0.9208	0.2058
	(100, 100)	95.8	0.7335	0.9134	0.1799
	(200, 150)	95.1	0.7631	0.8964	0.1333
(0.9, 0.8)	(50, 50)	95.0	0.7009	0.9267	0.2258
	(75, 75)	96.0	0.7281	0.9161	0.1880
	(100, 100)	95.4	0.7441	0.9089	0.1648
	(200, 150)	95.7	0.7691	0.8916	0.1224

NOTE: : CP(%): coverage probability, LE: the average left endpoint, RE: the average right endpoint and AL: the average length of the interval.

Chapter 5

SUMMARY AND FUTURE WORK

5.1 Summary

In this thesis, we proposed a smoothed semi-empirical likelihood method to construct the confidence intervals for ROC curves with missing data in both populations. The approach is easy to understand, simple to implement, and efficient to compute. We first presented an imputation method to deal with missing completely at random data. Then it can be shown that the semi-empirical likelihood ratio under imputation is asymptotically distributed as a scaled chi-squared variable. The finite sample numerical performance of the inference is evaluated. All empirical coverage levels are close to the nominal levels 95%, even for small or moderate sample size. The coverage lengths of the confidence intervals are small. We may also notice that the result can be applied to complete data setting. Under this scenario, the response rates are $P_1 = P_2 = 1$. The asymptotic distribution of the semi-empirical likelihood statistic is found to be a χ^2_1 distribution. Thus the semi-empirical likelihood based confidence interval for Δ_q is constructed as

$$\{\Delta_q : -2\log \mathcal{R}(\Delta, \theta_{m,n}) \leq t_\alpha\}.$$

The main contribution of this thesis is that it extends the previous studies about

application of empirical likelihood ratio principle to the ROC curve analysis. Furthermore, the smoothed semi-empirical likelihood ratio statistic is established and the limiting distribution is proved. Moreover, we consider special but common settings such as semi-parametric distributions where missing data occur (MCAR) in this thesis.

5.2 Future Work

In the future, we may compare the semi-empirical likelihood method to the normal approximation method and perhaps other non-parametric ways, such as bootstrap percentile method. For example, Su *et al.* (2009) summarized the idea of bootstrap confidence intervals for the ROC curve; Liang and Zhou (2008) examined the normal approximation-based confidence intervals for censored ROC curves and established the asymptotic result. Following their ideas, we may develop and investigate the performance of those methods to semi-parametric setting with missing data. In addition, we may apply the proposed approach to a real data application when we have a good data set. Moreover, we may develop a better imputation method for the missing data instead of the ad hoc hot deck imputation. We can also perform a simulation study on the complete data setting. Extension of our approach to the comparison of ROC curves needs further investigation.

REFERENCES

Chen, J. and Rao, J. N. K. (2007) Asymptotic normality under two-phase sampling designs. *Statistica Sinica*, **17**, 1047–1064.

Chen, S. X. (1993) On the accuracy of empirical likelihood confidence regions for linear regression model. *Annals of the Institute of Statistical Mathematics*, **45**, 621–637.

— (1994) Empirical likelihood confidence intervals for linear regression coefficients. *Journal of Multivariate Analysis*, **49**, 24–40.

Chen, S. X. and Hall, P. (1993) Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, **21**, 1166–1181.

Chen, S. X. and Qin, J. (2003) Empirical likelihood-based confidence intervals for data with possible zero observations. *Statistics & Probability Letters*, **65**, 29–37.

Claeskens, G., Jing, B.-Y., Peng, L. and Zhou, W. (2003) Empirical likelihood confidence regions for comparison distributions and ROC curves. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, **31**, 173–190.

DiCiccio, T., Hall, P. and Romano, J. (1991) Empirical likelihood is Bartlett-correctable. *The Annals of Statistics*, **19**, 1053–1061.

Gastwirth, J. L. and Wang, J.-L. (1988) Control percentile test procedures for censored data. *Journal of Statistical Planning and Inference*, **18**, 267–276.

Goddard, M. J. and Hinberg, I. (1990) Receiver operator characteristic (ROC) curves and non-normal data: An empirical study. *Statistics in Medicine*, **9**, 325–337.

Hall, P. and La Scala, B. (1990) Methodology and algorithms of empirical likelihood. *International Statistical Review*, **58**, 109–127.

Hartley, H. O. and Rao, J. N. K. (1968) A new estimation theory for sample surveys. *Biometrika*, **55**, 547–557.

Hollander, M. and Korwar, R. M. (1982) Nonparametric Bayesian estimation of the horizontal distance between two populations. In *Nonparametric Statistical Inference (in two volumes)* (eds. B. V. Gnedenko, M. L. Puri and I. Vincze), 409–416. Elsevier/North-Holland [Elsevier Science Publishing Co., New York; North-Holland Publishing Co., Amsterdam].

Hsieh, F. and Turnbull, B. W. (1996) Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, **24**, 25–40.

Li, G., Tiwari, R. C. and Wells, M. T. (1996) Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers. *Journal of the American Statistical Association*, **91**, 689–698.

— (1999) Semiparametric inference for a quantile comparison function with applications to receiver operating characteristic curves. *Biometrika*, **86**, 487–502.

Liang, H. and Zhou, Y. (2008) Semiparametric Inference for ROC Curves with Censoring. *Scandinavian Journal of Statistics*, **35**, 212–227.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*. John Wiley & Sons, 2nd edn.

Lloyd, C. J. (1998) Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*, **93**, 1356–1364.

Owen, A. (1990) Empirical likelihood ratio confidence regions. *The Annals of Statistics*, **18**, 90–120.

Owen, A. B. (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

— (2001) *Empirical Likelihood*. Chapman & Hall Ltd.

Pepe, M. S. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: Oxford University Press.

Qin, J. (1994) Semi-empirical likelihood ratio confidence intervals for the difference of two sample means. *Annals of the Institute of Statistical Mathematics*, **46**, 117–126.

— (1999) Empirical likelihood ratio based confidence intervals for mixture proportions. *The Annals of Statistics*, **27**, 1368–1384.

Qin, J. and Lawless, J. (1994) Empirical likelihood and general estimating equations. *The Annals of Statistics*, **22**, 300–325.

Qin, Y. and Zhang, J. (2009) Semi-empirical likelihood confidence intervals for the differences of quantiles with missing data. *Acta Mathematica Sinica, English Series*, **25**, 845–854.

Qin, Y. S. (1997) Semi-parametric likelihood ratio confidence intervals for various differences of two populations. *Statistics & Probability Letters*, **33**, 135–143.

Ren, H., Zhou, X.-H. and Liang, H. (2004) A flexible method for estimating the ROC curve. *Journal of Applied Statistics*, **31**, 773–784.

Shapiro, D. E. (1999) The interpretation of diagnostic tests. *Statistical Methods in Medical Research*, **8**, 113–134.

Su, H., Qin, Y. and Liang, H. (2009) Empirical Likelihood-Based Confidence Interval of ROC Curves. *Statistics in Biopharmaceutical Research*, **1**, 407–414.

Thomas, D. R. and Grunkemeier, G. L. (1975) Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, **70**, 865–871.

Tosteson, A. A. N. and Begg, C. B. (1988) A general regression methodology for roc curve estimation. *Medical Decision Making*, **8**, 204–215.

Wang, Q. and Rao, J. N. K. (2002) Empirical likelihood-based inference under for missing response data. *The Annals of Statistics*, **30**, 896–924.

Zhou, W. and Jing, B.-Y. (2003) Smoothed empirical likelihood confidence intervals for the difference of quantiles. *Statistica Sinica*, **13**, 83–95.

Zhou, X.-H., McClish, D. K. and Obuchowski, N. A. (2002) *Statistical Methods in Diagnostic Medicine*. Wiley.

Zhou, Y. and Liang, H. (2005) Empirical-likelihood-based semiparametric inference for the treatment effect in the two-sample problem with censoring. *Biometrika*, **92**, 271–282.

Zou, K. H., Hall, W. J. and Shapiro, D. E. (1997) Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, **16**, 2143–2156.

Zweig, M. and Campbell, G. (1993) Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, **39**, 561–577.

APPENDICES

Appendix A: Lemmas and Proofs

Following the idea of Qin and Zhang (2009), we have the following lemmas to prove the main result. Lemma A.1 is from Chen and Rao (2007).

Lemma A.1. *Let U_n, V_n be two sequences of random variables and $\mathcal{B}_n$ be a σ-algebra. Assume that*

1. *There exists $\sigma_{1n} > 0$ such that*

$$\sigma_{1n}^{-1} V_n \xrightarrow{d} N(0,1)$$

as $n \to \infty$, and V_n is $\mathcal{B}_n$ measurable;

2. *$E[U_n|\mathcal{B}_n] = 0$ and $Var(U_n|\mathcal{B}_n) = \sigma_{2n}^2$ such that*

$$\sup_t |P(\sigma_{2n}^{-1} U_n \leq t|\mathcal{B}_n|) - \Phi(t)| = o_p(1),$$

where $\Phi(.)$ is the distribution function of the standard normal random variables.

3. *$\gamma_n^2 = \sigma_{1n}^2/\sigma_{2n}^2 = \gamma^2 + o_p(1)$. Then, as $n \to \infty$,*

$$\frac{U_n + V_n}{\sqrt{\sigma_{1n}^2 + \sigma_{2n}^2}} \xrightarrow{d} N(0,1).$$

To prove the main result, we need some additional lemmas.

Lemma A.2. *Under the conditions of Theorem 3.1, as* $m, n \to \infty$,

$$\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \omega(x_{I,i}, \theta_0, \Delta_q) \xrightarrow{d} N(0, \sigma_1^2)$$

and

$$\frac{1}{m} \sum_{i=1}^{m} \omega^2(x_{I,i}, \theta_0, \Delta_q) = \beta_1^2 + o_p(1),$$

where

$$\sigma_1^2 = (1 - P_1 + P_1^{-1})\Delta_q(1 - \Delta_q), \quad \beta_1^2 = \Delta_q(1 - \Delta_q).$$

Proof. Let $\bar{\omega}_r = \frac{1}{r_x} \sum_{i \in s_{rx}} \omega(x_i, \theta_0, \Delta_q)$, and $\mathcal{B}_m = \sigma\{(\delta_{xi}, x_i), i = 1, ..., m\}$. Then

$$E(\omega(x_i^*, \theta_0, \Delta_q)|\mathcal{B}_m) = \bar{\omega}_r, \mathrm{Var}(\omega(x_i^*, \theta_0, \Delta_q)|\mathcal{B}_m) = \frac{1}{r_x} \sum_{i \in s_{rx}} \{\omega(x_i, \theta_0, \Delta_q) - \bar{\omega}_r\}^2.$$

It follows that

$$\begin{aligned} \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \omega(x_{I,i}, \theta_0, \Delta_q) &= \sqrt{m}\bar{\omega}_r + \frac{1}{\sqrt{m}} \sum_{i \in s_{mx}} [\omega(x_i^*, \theta_0, \Delta_q) - E\{\omega(x_i^*, \theta_0, \Delta_q)|\mathcal{B}_m\}] \\ &=: V_m + U_m, \end{aligned}$$

where V_m is $\mathcal{B}_m$ measurable, and

$$V_m = \sqrt{m}\frac{1}{r_x} \sum_{i \in s_{rx}} \{\omega(x_i, \theta_0, \Delta_q) - E\omega(x_i, \theta_0, \Delta_q)\} + \sqrt{m}E\omega(x_i, \theta_0, \Delta_q).$$

It can be shown that $E\omega(x_i, \theta_0, \Delta_q) = O(h^{t_0})$. Thus from Assumption (iii) and (v), $\sqrt{m}E\omega(x_i, \theta_0, \Delta_q) = o(1)$. Combining with the MCAR assumption and the Central Limit Theorem,

$$V_m \xrightarrow{d} N(0, P_1^{-1}\Delta_q(1 - \Delta_q)).$$

From Berry-Esseen's Central Limit Theorem for independent random variables, we have $\sup_t |P(\sigma_{2m}^{-1} U_m \le t|\mathcal{B}_m) - \Phi(t)| = o_p(1)$, where $\sigma_{2m}^2 = (1-P_1)E\omega^2(x_i,\theta_0,\Delta_q) = (1-P_1)\Delta_q(1-\Delta_q)$. Hence, from Lemma A1, we have

$$\frac{1}{\sqrt{m}}\sum_i \omega(x_{I,i},\theta_0,\Delta_q) \xrightarrow{d} N(0,\sigma_1^2).$$

On the other hand, denote the conditional probability given $\mathcal{B}_m$ as P^*. Then by the law of large numbers and MCAR assumption,

$$\frac{1}{\sqrt{m_x}}\sum_{i\in s_{mx}} \omega^2(x_i^*,\theta_0,\Delta_q) = \frac{1}{r_x}\sum_{i\in s_{rx}} \omega^2(x_i,\theta_0,\Delta_q) + o_{p^*}(1) = E\omega^2(x,\theta_0,\Delta_q) + o_p(1).$$

It follows that

$$\begin{aligned}
\frac{1}{\sqrt{m_x}}\sum_{i=1}^m \omega^2(x_{I,i},\theta_0,\Delta_q) &= \frac{1}{m}\sum_{i=1}^m \{\delta_{xi}\omega^2(x_i,\theta_0,\Delta_q) + (1-\delta_{xi})\omega^2(x_i^*,\theta_0,\Delta_q)\} \\
&= P_1 E\omega^2(x_i,\theta_0,\Delta_q) + o_p(1) + \frac{m_x}{m}\frac{1}{m_x}\sum_{i\in s_{mx}} \omega^2(x_i^*,\theta_0,\Delta_q) \\
&= P_1 E\omega^2(x,\theta_0,\Delta_q) + o_p(1) + (1-P_1)E\omega^2(x,\theta_0,\Delta_q) + o_p(1) \\
&= E\omega^2(x,\theta_0,\Delta_q) + o_p(1) \\
&= \Delta_q(1-\Delta_q) + o_p(1).
\end{aligned}$$

□

Lemma A.3. *Suppose that $1/3 < \eta < 1/2$ and the conditions of Theorem 3.1 are satisfied. Then as $m, n \to \infty$,*

$$\lambda(\theta) = \mathrm{O}_P(n^{-\eta}h^{-1} + h^{t_0}),$$

uniformly about $\theta \in \{\theta : |\theta - \theta_0| \le cn^{-\eta}\}$, where c is some positive constant.

Proof. The proof of Lemma A.3 is omitted, which is in a similar way as the proof of Lemma 4.3 in Qin and Zhang (2009). □

Lemma A.4. *Suppose that* $1/3 < \eta < 1/2$ *and the conditions of Theorem 3.1 are satisfied. Then with probability tending to 1, there exists a root* $\hat{\theta}_{EL}$ *of estimating equation (3.6) such that, as* $m, n \to \infty$,

$$|\hat{\theta}_{EL} - \theta_0| = \mathrm{O}_p(n^{-\eta}),$$

and $\mathcal{R}(\Delta_q, \theta)$ *attains its local maximum value at* $\hat{\theta}_{EL}$.

Proof. The proof of Lemma A.4 is omitted, which is essentially in a similar way as the proof of Lemma 4.4 in Qin and Zhang (2009). □

Lemma A.5. *Suppose that the conditions of Theorem 3.1 are satisfied, and* $\hat{\theta}_{EL}$ *is as that in Lemma A4. Then, as* $m, n \to \infty$,

$$\sqrt{m}\begin{pmatrix} \hat{\theta}_{EL} - \theta_0 \\ \lambda(\hat{\theta}_{EL}) \end{pmatrix} \xrightarrow{d} N(0, \Sigma),$$

where

$$\Sigma = \frac{1}{c_1^2}\begin{pmatrix} \beta_0^2\sigma_1^2 + kP_2\{\Delta_q(1-\Delta_q)\}^2 I(\theta_0) & kP_2\beta_0 I(\theta_0)\{\Delta_q(1-\Delta_q) - \sigma_1^2\} \\ kP_2\beta_0 I(\theta_0)\{\Delta_q(1-\Delta_q) - \sigma_1^2\} & kP_2 I(\theta_0)\{\beta_0^2 + kP_2\sigma_1^2 I(\theta_0)\} \end{pmatrix},$$

$\beta_0 = \alpha(\theta_0) f(F^{-1}(1-\Delta_q)), \quad \sigma_1^2 = (1 - P_1 + P_1^{-1})\Delta_q(1-\Delta_q), \quad c_1 = \beta_0^2 + kP_2\Delta_q(1-\Delta_q)I(\theta_0)$.

Proof. The proof of Lemma A.5 follows the idea of the proof of Lemma 4.5 in Qin

and Zhang (2009). Let $\lambda = \lambda(\theta)$, $\lambda_{EL} = \lambda(\hat{\theta}_{EL})$, and

$$\begin{aligned} Q_{1,m,n}(\theta,\lambda) &= \frac{1}{m}\sum_{i=1}^{m}\frac{\omega(x_{I,i},\theta,\Delta_q)}{1+\lambda(\theta)\,\omega(x_{I,i},\theta,\Delta_q)}, \\ Q_{2,m,n}(\theta,\lambda) &= \frac{\lambda}{m}\sum_{i=1}^{m}\frac{h^{-1}K[\{G_\theta^{-1}(1-q)-x_{I,i}\}/h]}{1+\lambda\,\omega(x_{I,i},\theta,\Delta)}\,\alpha(\theta) - \frac{1}{m}\sum_{j=1}^{n}\delta_{yj}\frac{\partial \log g_\theta(y_j)}{\partial\theta}. \end{aligned}$$

From Lemma A.4, we have

$$Q_{i,m,n}(\hat{\theta}_{EL},\lambda_{EL}) = 0, \quad i = 1,2.$$

Based on Taylor expansion, Lemmmas A.3 and A.4, we have

$$\begin{aligned} 0 &= Q_{i,m,n}(\hat{\theta}_{EL},\lambda_{EL}) = Q_{i,m,n}(\theta_0,0) + \frac{\partial Q_{i,m,n}(\theta_0,0)}{\partial\theta}(\hat{\theta}_{EL}-\theta_0) \\ &\quad + \frac{\partial Q_{i,m,n}(\theta_0,0)}{\partial\lambda}\lambda_{EL} + o_p(\xi_n), \quad i = 1,2, \end{aligned}$$

where $\xi_n = |\hat{\theta}_{EL}-\theta_0| + |\lambda_{EL}|$. Thus,

$$Q_{i,m,n}(\theta_0,0) + \frac{\partial Q_{i,m,n}(\theta_0,0)}{\partial\theta}(\hat{\theta}_{EL}-\theta_0) + \frac{\partial Q_{i,m,n}(\theta_0,0)}{\partial\lambda}\lambda_{EL} = o_p(\xi_n), \quad i = 1,2.$$

Similar to the proof of Lemma A.2, it can be shown that

$$\begin{aligned} \frac{\partial Q_{1,m,n}(\theta_0,0)}{\partial\theta} &= \alpha(\theta_0) f\{F^{-1}(1-\Delta_q)\} + o_p(1), \\ \frac{\partial Q_{1,m,n}(\theta_0,0)}{\partial\lambda} &= -\Delta_q(1-\Delta_q)\} + o_p(1), \\ \frac{\partial Q_{2,m,n}(\theta_0,0)}{\partial\theta} &= kP_2 I(\theta_0) + o_p(1), \\ \frac{\partial Q_{2,m,n}(\theta_0,0)}{\partial\lambda} &= \alpha(\theta_0) f\{F^{-1}(1-\Delta_q)\} + o_p(1). \end{aligned}$$

Hence,

$$\begin{pmatrix} \hat{\theta}_{EL} - \theta_0 \\ \lambda(\hat{\theta}_{EL}) \end{pmatrix} = S^{-1} \begin{pmatrix} -Q_{1,m,n}(\theta_0, 0) \\ -Q_{2,m,n}(\theta_0, 0) \end{pmatrix} + o_p(\xi_n),$$

where

$$S = \begin{pmatrix} \beta_0 & -\Delta_q(1-\Delta_q) \\ kP_2I(\theta_0) & \beta_0 \end{pmatrix}.$$

Then, Lemma A.2 and the central limit theorem lead to

$$\sqrt{m} \begin{pmatrix} Q_{1,m,n}(\theta_0, 0) \\ Q_{2,m,n}(\theta_0, 0) \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & kP_2I(\theta_0) \end{pmatrix}\right).$$

It follows from assumption (viii) that

$$\sqrt{n}\, Q_{j,m,n}(\theta_0, 0) = O_p(1), \quad j = 1, 2,$$

and thus, $\xi_n = O_p(n^{-1/2})$. This complete the proof of Lemma A.5. □

The following Lemma can be proved using the method in the proof of Lemma A.2. Denote $\bar{\omega}_j(\theta) = m^{-1} \sum_{i=1}^{m} \omega^j(x_{I,i}, \theta, \Delta)$ for $j = 1, 2$.

Lemma A.6. *Under the conditions of Theorem 3.1, as $m, n \to \infty$,*

$$\bar{\omega}_2(\hat{\theta}_{EL}) = \Delta_q(1-\Delta_q) + o_p(1), \quad \partial^2 l_2(\hat{\theta}_{EL})/\partial^2\theta = -kP_2I(\theta_0) + o_p(1),$$

$$\partial^2 l_2(\hat{\theta})/\partial^2\theta = -kP_2I(\theta_0) + o_p(1),$$

where $\bar{\omega}_2(\theta)$ and $l_2(\theta)$ are defined in the proof of Lemma A.4 and Assumption (xi), respectively.

Proof of Theorem 3.1. Use the notations of $\bar{\omega}_j(\theta)$ in the proof of Lemma A.4. From

Taylor expansion, it follows that

$$\begin{aligned} & -2\log\mathcal{R}(\Delta,\hat{\theta}_{EL}) \qquad (1)\\ = \;& 2\sum_{i=1}^{m}\log\{1+\lambda(\hat{\theta}_{EL})\omega(x_{I,i},\theta,\Delta_q)\}-2m\{l_2(\hat{\theta}_{EL})-l_2(\hat{\theta})\}\\ = \;& 2m\lambda(\hat{\theta}_{EL})\bar{\omega}_1(\hat{\theta}_{EL})-m\lambda^2(\hat{\theta}_{EL})\bar{\omega}_2(\hat{\theta}_{EL})+2m\{\partial l_2(\hat{\theta}_{EL})/\partial\theta\}(\hat{\theta}-\hat{\theta}_{EL})\\ + \;& m\{\partial^2 l_2(\hat{\theta}_{EL})/\partial\theta^2\}(\hat{\theta}-\hat{\theta}_{EL})^2+o_p(1). \end{aligned}$$

Equation (3.5) gives that $\bar{\omega}_1(\hat{\theta}_{EL})=\lambda(\hat{\theta}_{EL})\bar{\omega}_2(\hat{\theta}_{EL})+o_p(n^{-1/2})$.

From $\partial l_2(\hat{\theta})/\partial\theta=0$ and Taylor expansion, we have

$$\partial l_2(\hat{\theta}_{EL})/\partial\theta=\{\partial^2 l_2(\hat{\theta})/\partial\theta^2\}(\hat{\theta}_{EL}-\hat{\theta})+o_p(n^{-1/2}).$$

Equation (3.6) leads to

$$\lambda(\hat{\theta}_{EL})\sum_{i=1}^{m}\frac{h^{-1}K[\{G^{-1}_{\hat{\theta}_{EL}}(1-q)-x_{I,i}\}/h]}{1+\lambda(\hat{\theta}_{EL})\omega(x_{I,i},\hat{\theta}_{EL},\Delta)}\,\alpha(\hat{\theta}_{EL})=\sum_{j=1}^{n}\delta_{yj}\frac{\partial\log g_{\hat{\theta}_{EL}(y_j)}}{\partial\theta}.$$

Thus combining with Lemma A.3, we have

$$\partial l_2(\hat{\theta}_{EL})/\partial\theta=\lambda(\hat{\theta}_{EL})\alpha(\theta_0)f(F^{-1}(1-\Delta_q))+o_p(n^{-1/2}).$$

Therefore,

$$\hat{\theta}_{EL}-\hat{\theta}=\lambda(\hat{\theta}_{EL})\alpha(\theta_0)f(F^{-1}(1-\Delta_q))\{\partial^2 l_2\hat{\theta}/\partial\theta^2\}^{-1}+o_p(n^{-1/2}).$$

From the above derivations and Lemma A.5 and A.6, we have Theorem 3.1 proved. □

Appendix B: R Code for Simulation

```
############################################################
#Simulation Studies
#Semi-empirical-likelihood-ROC-with-missing-data
#written by Xiaoxia Liu
#05-20-2010
############################################################


###################### Step 1 ##########################
#Generate random samples with missing values
### set the random number generating method
RNGkind(kind="default", normal.kind="default")
###-----------------------------------------------
#random seed
set.seed(520)
library(BB)

begin.time<-proc.time()
m=100 #sample size of x
n=100 #sample size of y
p1=0.9 #response rate of x
p2=0.8 #response rate of y
N=1000 # number of repetition
sim=1

###-----------------------------------------------
###Create vectors and matrix to store the results
theta.est.vec=rep(0, N)
lambda.est.vec=rep(0, N)
lower.bound.vec=rep(0, N)
upper.bound.vec=rep(0, N)
ci.length.vec=rep(0, N)
coverage.prob=rep(0, N)
message1=rep(0,N)
message2=rep(0,N)
message3=rep(0,N)
iter1.vec=rep(0,N)
iter2.vec=rep(0,N)
iter3.vec=rep(0,N)
```

```
temp2.mat=matrix(0, nrow=N, ncol=3)
temp3.mat=matrix(0, nrow=N, ncol=3)


#begin iteration
while (sim <=N){
###################### Step 2 ##########################
delta.x=as.numeric(runif(m)<=p1) #missing indictor
delta.y=as.numeric(runif(n)<=p2)
rx=sum(delta.x)
ry=sum(delta.y)
mx=m-rx  #sets of nonrespondents
my=n-ry
x0=rnorm(m, mean=1, sd=1) #initial sample of complete data
y0=rnorm(n, mean=0, sd=1)
xr=x0[delta.x==1]  #sets of respondents
yr=y0[delta.y==1]
hot.deck=runif(mx)#generate unif prob
x1=xr[ceiling(rx*hot.deck)] #hot deck imputation
#x1=sample(xr,mx, replace =T) #hot deck imputation
theta.mle=mean(yr)
y1=rnorm(my, mean=theta.mle, sd=1)
x=x0*delta.x
#'complete' data after imputation
#(keep the original order of the observations)
x[delta.x==0]=x1
y=y0*delta.y
#'complete' data after imputation
#(keep the original order of the observations)
y[delta.y==0]=y1


###################### Step 3 ##########################
#semi-EL based confidence interval for delta
k.fun=function(u){exp(-u^2/2)/sqrt(2*pi)}  #define kernel K

hn=(m^(-1/3))*(3/2)  #bandwidth

alpha=0.05 #correct coverage probability

q0=0.3 #q-th quantile
```

```
delta0=1 - pnorm(qnorm((1-q0), mean=0, sd=1), mean=1, sd=1)#true delta

# empirical estimation of delta0
delta.est= 1- sum(xr <= qnorm((1-q0), mean=0, sd=1))/length(xr)

H.fun=function(t) {pnorm(t/hn, mean = 0, sd = 1)} #H(t)

omega.fun=function(xi, theta, Delta){
 H.fun(qnorm((1-q0), mean=theta, sd=1)-xi)-(1-Delta)
} #omega

#alpha function under normal distribution
#which equals 1
alpha.fun=function(theta) {1}


k=n/m
beta0.est=function(theta){
 sum(k.fun((qnorm((1-q0), mean=theta, sd=1) - x)/hn))
 *alpha.fun(theta) /(m*hn)
}  #beta0

sigma1.est=function(delta){
 sqrt((1-sum(delta.x)/m + m/sum(delta.x))*delta*(1-delta))
}  #sigma square

c1=function(delta, theta){
 beta0.est(theta)^2 + k*(sum(delta.y)/n)*delta*(1-delta)
}#c1

rho0fun.est=function(delta, theta) {
 c1(delta, theta)/(k*(sum(delta.y)/n)
 *sigma1.est(delta)^2 + beta0.est(theta)^2)
} # the function a0

#loglik function
Rfun=function(delta, lambda, theta) {
 -sum(log(1+ lambda*omega.fun(x, theta, delta)))
 + sum(delta.y*(-0.5*(y-theta)^2))
 - sum(delta.y*(-0.5*(y-theta.mle)^2))
}
```

```
Rfun2 = function(delta, lambda, theta) {
 -2*rho0fun.est(delta, theta)* Rfun(delta, lambda, theta)
 - qchisq(p=1-alpha, df=1)
}

est.fun2=function(para){
#para[1]: lambda; para[2]: delta; para[3]: theta
g=rep(NA, length(para))
g[1] = sum(omega.fun(x, para[3], para[2])
/(1 + para[1]*omega.fun(x, para[3], para[2])))
###
g[2] = -2* rho0fun.est(para[2], para[3])
* Rfun(para[2], para[1], para[3])
- qchisq(p=1-alpha, df=1)
###
g[3] = para[1] *sum((1/hn)
*k.fun((qnorm((1-q0), mean=para[3], sd=1)-x)/hn)/
(1+para[1]*omega.fun(x,para[3],para[2])))
 -sum(delta.y*(y-para[3]))
###
g
}

# Now, we solve the equation using BBsolve.
#starting value for lower bound
v1=c(-0.3, -0.35, -0.4, -0.2, -0.25, -0.5, -0.45, -0.6)
v2=c(0.5, 0.45, 0.55, 0.4, 0.6)
v3=c(0.2, 0.1)

para2.mat = as.matrix(expand.grid(v1, v2, v3))
iter2=1
temp.mess2="Unsuccessful convergence"

while(iter2<=dim(para2.mat)[1] & temp.mess2!="Successful convergence"){
 para2=para2.mat[iter2,]
 eqn2=BBsolve(par=para2, fn=est.fun2, method=1)
 temp2=eqn2$par
 temp.mess2=eqn2$message
 iter2=iter2+1
}

iter2.vec[sim]=iter2
```

```
temp2.mat[sim,]=temp2
message2[sim]=(temp.mess2 == "Successful convergence")*1

#starting value for upper bound
v4=c(0.4, 0.5, 0.3, 0.45, 0.35, 0.6, 0.7, 0.8)
v5=c(0.75, 0.7, 0.8, 0.85, 0.9, 0.95)
v6=c(-0.1, -0.2, -0.3)

para3.mat = as.matrix(expand.grid(v4, v5, v6))
iter3=1
temp.mess3="Unsuccessful convergence"

while(iter3<=dim(para3.mat)[1] & temp.mess3!="Successful convergence"){
 para3=para3.mat[iter3,]
 eqn3=BBsolve(par=para3, fn=est.fun2)
 temp3=eqn3$par
 temp.mess3=eqn3$message
 iter3=iter3+1
}

iter3.vec[sim]=iter3
temp3.mat[sim,]=temp3
message3[sim]=(temp.mess3 == "Successful convergence")*1


lower.bound.vec[sim] = min(temp2[2], temp3[2])
upper.bound.vec[sim] = max(temp2[2], temp3[2])
ci.length.vec[sim] = upper.bound.vec[sim]-lower.bound.vec[sim]
coverage.prob[sim] = (delta0 >= lower.bound.vec[sim])
* (delta0 <= upper.bound.vec[sim])

cat("iteration = ", sim, "\n")

sim = sim +1
}


end.time<-proc.time()-begin.time

save.image("/D:/simu/setup1/emp-lkhd-p1-09-p2-08-m100-n100-q-03.RData")
```

```
################## summary of results ##################
#true delta_q
delta0

#CP
sum(coverage.prob)/N*100

#LE
#mean(lower.bound.vec)
round(mean(lower.bound.vec), digits=4)

#RE
#mean(upper.bound.vec)
round(mean(upper.bound.vec), digits=4)

#AL
#mean(ci.length.vec)
round(mean(ci.length.vec), digits=4)

#p1, p2
c(p1, p2)

#m, n
c(m, n)

#########################################################
################### end of my R code ####################
```