8-11-2020

Some Novel Interval Estimation Methods

Xinjie Hu

SOME NOVEL INTERVAL ESTIMATION METHODS

by

XINJIE HU

Under the Direction of Gengsheng Qin, PhD

ABSTRACT

In medical diagnostic studies, the Youden index is a summary measure widely used in the evaluation of the diagnostic accuracy of a medical test. When covariates are not considered, the diagnostic accuracy of the test can be biased or misleading. By incorporating information from covariates, we propose and compare various confidence intervals for the covariate-adjusted Youden index and its optimal cut-off point. In ROC analysis, the area under the ROC curve (AUC) is a popular one number summary index of the discriminatory accuracy of a diagnostic test. Adjustment for covariate effects can greatly improve the diagnostic accuracy of a test for individual patient. AUC Regression is widely used to evaluate the effects of the covariates on the diagnostic accuracy. Using side information provided by the influence function, empirical likelihood methods are proposed for inferences of AUC in the presence of covariates. For parameters in the AUC regression model, it is shown that the asymptotic distribution of the influence function-based empirical log-likelihood ratio statistic

is a standard chi-square distribution. Hence, confidence regions for the regression parameters can be easily obtained without any variance estimates.

The latter half of this dissertation focuses on empirical likelihood (EL) based interval estimation methods for correlation coefficient (CC) and coefficient of variation (CV). Under normal distribution assumptions, there are many types of confident intervals for CC or CV, such as the GPQ-based 'exact' interval, the Z transformation-based interval, and maximum likelihood-based intervals. However, the exact method is computationally cumbersome, and approximation methods can't be applied when the underlying distribution is unknown. Therefore, we propose influence function-based empirical likelihood intervals for CC and CV. Extensive simulation studies are conducted to evaluate the finite sample performances of the proposed EL-based intervals in terms of coverage probability. Finally, we illustrate the proposed methods with real examples.

Key words: AUC Regression, Bootstrap, Coefficient of Variation, Correlation Coefficient, Covariates, Empirical Likelihood, Fisher Z-transformation, Generalized Pivotal Quantity, Influence function, Jackknife, ROC Curve, Youden Index.

SOME NOVEL INTERVAL ESTIMATION METHODS

by

XINJIE HU

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

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2020

SOME NOVEL INTERVAL ESTIMATION METHODS

by

XINJIE HU

Committee Chair: Gengsheng Qin

Committee: Jun Kong

Jing Zhang

Ruiyan Luo

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2020

DEDICATION

To my parents, my advisor, and all my college friends.

ACKNOWLEDGEMENTS

I'm profoundly indebted to my thesis advisor Dr. Gengsheng Qin. I am deeply grateful that Dr. Qin believed in my potential and accepted me as his student. During the four years I worked with him, I have come to know Dr. Qin as a brilliant professional with deep and broad knowledge, accurate sense of intuition and infinite enthusiasm for inquisition and creation. Dr. Qin taught me a great deal of statistics, probability and mathematics knowledge with patience and generosity. I am grateful for his valuable time offered to me, for the inspiring discussion on research and for the friendly encouragement and support. He helped me transition from an dependent learner to an independent researcher and instructor. The most invaluable qualities Dr. Qin instilled upon me are the ability to identify a meaningful statistical problem and the tenacity of cracking the problem even after a series of frustrations and failures. Rather than studying guidance, Dr. Qin also affectionately offered me valuable suggestions for my future career search based on his understanding of my characteristics and personalities. My appreciations to Dr. Qin are ceaseless. He puts great expectation and abundant trusts on my ability and potential, which encourages me to take the challenges for problem solving even after a series of failures. His passion, optimism and diligence towards research are extremely contagious, which were a great source of inspiration during the course of my thesis research, and will continue to inspire me to be a better researcher.

I would like to express my sincere gratitude to a group of great professors and researchers both inside and outside GSU for the advice and support I received from them: Dr. Yi Jiang, Dr. Zhongshan Li, Dr. Xin Qi, Dr. Xingxing Yu, and Dr. Yu Feng. I would like to thank my thesis committee, Dr. Jun Kong, Dr. Jing Zhang and Dr. Ruiyan Luo for providing invaluable advices and comments regarding the dissertation material and the presentation of this thesis. My heartfelt appreciation goes to Chenxue Li, Jinyuan Chen, Kaihua Cai, Baoying Yang, Binghuan Wang and Haochuan Zhou for their enthusiastic discussions with

me, sharing enlightening and novel ideas and unique experience regarding the theory behind topics.

I will miss the time I spent with my officemate and fellow PhD students, Yan Hai, Guanhao Wei, Hao Chen, Guangming Jin, Bin Zhang, Haiqi Wang, Yuyin Shi and Bin Liu. I shared the same office with Yan Hai and the fact that we work on different topics did not prevent us from learning from each other. As more senior members of graduate students in the department, they were so generous in sharing their perspective and experience. Yan Hai, Bin Liu and Guanhao Wei are my academic sister and brother, having the same passion against empirical likelihood methods as I do. I thank them for the stimulating discussion, camaraderie and caring they provided. My life wouldn't be such colorful without my friends and colleagues in GSU. It's great pleasure that working in the group directed by Dr. Qin. We have tremendous great time to discuss our research problems and share opinions with each other. I cherish those moments particularly when we conduct our group meeting each Friday afternoon to present our research progress of one week. It stimulate us transform from a ideas writer to an interpreter and instructor. We have same motivation towards our goals and we are proud of each other even for the small achievement. It's my fortunate to meet all of you in GSU during the past five years.

I wish to thank the amazing group of staffs and professors in the mathmatics and statistics department,Dr. Alexandra Smirnova, Dr. Changyong Zhong and Dr. Mark Grinshpon, Dr. Yichuan Zhao, Sandra Ahuama-Jonas, Beth Conner and Earnestine Collier, who worked so hard to take good care of us and provided generous help for our research and teaching duties. My special thanks go to Beth for her warm and smiling friendship.

I would not have completed the thesis without the unconditional love and support from my wonderful parents. Words cannot express how grateful I am to my parents for all of the sacrifices that they've made on my behalf. They always believe in me, provide the best

assistance for me and stand ready to advice and comfort in time of need. So, thank you, Mom and Dad, for being the most loving and supportive family one could every wish for. My success today is credited to their understanding and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

- GSU - Georgia State University
- YI - Youden Index
- ROC - Receiver Operating Characteristic
- AUC - Area under ROC curve
- VUS - Volume under ROC surface
- GPQ - Generalized Pivotal Quantities
- HAC - Hybrid Agresti-Coull
- HWS - Hybrid Wilson-Score
- HBCA - Hybrid Bias Correction and Acceleration
- ACNA - Agresti-Coull Bootstrap-based
- BCa - Bias Correction and Acceleration
- MOVER - Method Of Variance Estimates Recovery
- EL - Empirical Likelihood
- JEL - Jackknife Empirical Likelihood
- CC - Correlation Coefficient
- CV - Coefficient of Variation
- NAI - Normal Approximation Interval
- NAII - Hybrid Maximum Likelihood and Bootstrap Interval

- PEL - Plug-in Empirical Likelihood
- IFEL - Influence Function Empirical Likelihood
- MLE - Maximum Likelihood Estimator
- BP - Bootstrap Percentile
- BPEL - Bootstrap and Empirical Likelihood
- c.d.f - Cumulative Distribution Function

PART 1

INTRODUCTION

Since the normality assumption for the underlying distribution of population may not be easily guaranteed, and in many cases, the underlying distribution is not normal or unknown, the parametric methods inferences become quite involved. In this dissertation, we focus on developing the non-parametric interval estimation methods for the parameter of interests. In part 2, we apply the MOVER method and binomial proportion confidence intervals for the covariate-adjusted YI and its optimal cut-off point with/without normal error assumptions. It is shown that the asymptotic distribution of the influence function-based empirical log-likelihood ratio statistic is a standard chi-square distribution. Hence, confidence regions for the regression parameters can be easily obtained without any variance estimates. In part 3-5, we focus on developing empirical likelihood (EL) based non-parametric methods, particularly we propose the influence function-based empirical likelihood (IFEL) based interval estimation method, to construct confidence intervals/regions for the parameters in the AUC regression model along with the correlation coefficient (CC) and the coefficient of variation (CV) as well. Extensive simulation studies are conducted to evaluate the finite sample performances of the proposed EL-based intervals in terms of coverage probability. Finally, we illustrate the proposed methods with real examples.

1.1 Statistical Evaluation of Diagnostic Tests

In medical diagnostic studies, a continuous-scale test is usually applied to distinguish diseased and non-diseased populations from each other. Without loss of generality, we assume that higher test values indicate higher probability of the disease. Assume that X and Y are values of a continuous-scale test for non-diseased and diseased subjects, respectively. A subject is classified as diseased (positive) if the subject's test value is greater than or equal

to a chosen threshold value "c". Each threshold value "c" is associated with a probability of a true positive (TPR) result known as the sensitivity defined by $P(X < c)$ and a probability of a true negative result (TNR) known as specificity defined by $P(Y > c)$ respectively. The false negative rate (FNR) is defined as FNR = 1 - TPR and the false positive rate (FPR) is defined as FPR = 1 - TNR. The receiver operating characteristic (ROC) curve is the plot of sensitivity (q) versus 1-specificity ($1 - p$) for all possible threshold values. The ROC curve has been widely used to evaluate the performance of a test for discriminating diseased and non-diseased populations.

In ROC analysis, the Youden index (YI) [1] is a commonly used summary measure of the diagnostic accuracy of a test. The YI of the test is defined as max $\{p + q - 1\}$, in which the maximum is taken over all p's and q's on the ROC curve, or equivalently over all possible threshold values. The value for YI is between 0 and 1. Then the corresponding threshold value for YI is called as the optimal cut-off point.

1.2 Covariates Adjustment for the Youden Index

When researchers perform a diagnostic trial, some covariates, like characteristics of study subjects or operating conditions for the test, may affect test results by influencing the distributions of test measurements for diseased and/or non-diseased subjects, respectively. Without incorporating the covariates information, ROC analysis may be biased or misleading. Therefore, it is an intuitive way to incorporate covariates in ROC curve analysis in order to make use of the additional information. Pardo-Fernandez et al. [2] gave an excellent review on ROC curve analysis in the presence of covariates.

In order to evaluate the influence of covariates on the YI, some researchers have used induced-regression methods. They modeled the test/biomarker values through regression models in each population separately. Pepe [3] and Tosteson et al. [4] specified models for test results as a function of disease status and covariates. Smith and Thompson [5] proposed a parametric survival model for modeling the distribution of the screening test outcome as a function of true disease status and other confounding covariates. Zhou et al.

[6] extended the models proposed in Pepe by allowing for heteroscedasticity. Zheng and Heagerty [7] proposed a semi-parametric estimator for the conditional ROC curve, in which the distribution of the error terms is unknown and allowed to depend on the covariates, but, as in the previous articles, the effect of the covariates on the conditional means and variances is modeled parametrically. Recently, Rodríguez and Martínez [8] presented a Bayesian semi-parametric model, in which the error terms are assumed to be normally distributed, but non-parametric specifications of the conditional means and variances are allowed.

Faraggi [9] used a simple linear regression to model biomarker values from the diseased and non-diseased populations, and provided adjusted confidence intervals for the YI and the corresponding threshold value by using a bootstrap method. In section 2 of part 2, we consider similar linear regression models to those used in Faraggi [9], and compare the GPQ, BCa, Bootstrap methods for construction of confident interval for adjusted Youden index along with the optimal cutoff point. While linear regression models may be too simple to connect covariates and test values within each population, Yao et al. [10] proposed the use of heteroscedastic regression models for test results and provided a covariate-adjusted estimator for the AUC. Zhou and Qin [11] proposed nonparametric estimators for the covariate-adjusted YI under heteroscedastic regression models. In section 3 of part 2, we further extend Zhou and Qin's work, generalize the approach of Faraggi and propose various confidence intervals for the covariate-adjusted YI and its associated cut-off point under heteroscedastic regression models with/without normal error assumptions.

1.3 Empirical Likelihood Methods

Empirical likelihood (EL), introduced by Owen [12][13], is a powerful non-parametric method and its advantages over the normal approximation-based methods have been well-recognized (see, e.g., Hall and La Scala, [14]). There are a number of advantages to the EL method over normal approximation methods. For example, EL-based method allows for confidence intervals/regions construction without a variance estimator; in the EL method, instead of assuming a symmetric shape for the confidence intervals/regions, the shape is

automatically determined by the data. Over last two decades, EL has found wide applications in many areas such as in econometrics, medical studies and survey sampling. Readers are referred to Owen's [15] book and references therein. EL-based methods have been successfully applied to ROC analysis (Claeskens *et al.*,[16], Qin and Zhou, [17]).

In part 3, we propose two empirical likelihood based confidence intervals for inferences of AUC regression in the presence of covariates: one is an influence function-based empirical likelihood confidence interval (IFEL), the other is a Jackknife empirical likelihood confidence interval (JEL). The proposed new methods allow for confidence region construction without a variance estimator. We also construct confidence intervals for the covariate-adjusted AUC. Simulation studies are conducted to compare the relative performance of the proposed EL-based methods with the existing method in AUC regression. It shows that the proposed methods have better small sample performances than existing normal approximation-based confidence region in terms of coverage probability.

In part 4, we apply the influence function-based EL method to construct a confidence interval for CC and compare it with the existing estimation methods. In part 5, we examine the EL-based confident intervals including the plug-in empirical likelihood-based PEL interval, bootstrap-based BPEL interval, the Jackknife empirical likelihood-based JEL interval, and the proposed influence function-based IFEL interval, with the existing intervals including the Vangel's approximation-based interval, the bootstrap percentile (BP) interval and GPQ-based 'exact' confidence interval for CV under normal/non-normal underlying distribution assumptions.

PART 2

CONFIDENCE INTERVALS FOR THE YOUDEN INDEX AND ITS OPTIMAL CUT-OFF POINT IN THE PRESENCE OF COVARIATES

2.1 Introduction

In literature, two approaches have been used to model the relationship between test values and covariates. The first approach is to model the dependence of the ROC curve directly on the covariates (Pepe, Dodd and Pepe [18], Pepe and Cai). However this approach loses the connection with the cut-off value and does not allow the prediction of the sensitivity and specificity at a given cut-off point conditional on covariates. The second approach is to directly model the covariate effects on the test results and through the modeling process obtain the covariate-adjusted ROC curve and its related summary measures. Faraggi [9] used a simple linear regression to model biomarker values from the diseased and non-diseased populations, and provided adjusted confidence intervals for the YI and the corresponding threshold value by using a bootstrap method. In this part, we consider similar linear regression models to those used in Faraggi, and provide a Generalized Pivotal Quantity (GPQ) (see Tsui and Weerahandi, [19]) based method for constructing 'exact' confidence intervals for the covariate-adjusted YI and its associated optimal cut-off point. In literature, GPQ-based inferences have been applied to many problems. For example, Gamage et al. [20] constructed a generalized confidence region for the difference between two mean vectors; Lee and Lin [21] developed confidence intervals for the ratio of the means of two normal populations; Tian and Wilding [22] presented a generalized variable approach for confidence interval estimation of a common correlation coefficient from several independent samples drawn from bivariate normal populations; Tian [23] provided confidence intervals for the AUC with normal outcomes in linear models. Recently, Lai et al. [24] made use of a generalized approach to construct confidence intervals for the YI and its corresponding optimal cut-off point. Further details

on generalized confidence intervals can be found in Weerahandi [25][26][27].

While linear regression models may be too simple to connect covariates and test values within each population, Yao et al. [10] proposed the use of heteroscedastic regression models for test results and provided a covariate-adjusted estimator for the AUC. Zhou and Qin [11] proposed nonparametric estimators for the covariate-adjusted YI under heteroscedastic regression models. Inácio de Carvalho et al. [28] developed a nonparametric Bayesian covariate-adjusted estimation for the YI. In this part, we further extend Zhou and Qin's work and propose various confidence intervals for the covariate-adjusted YI and its associated cut-off point under heteroscedastic regression models.

In section 2.2, we incorporate information from covariates using induced linear regression models for test results. In section 2.3, under heteroscedastic regression models, we compare various confidence intervals including Wilson Score HWS confidence interval, Agresti-Coull HAC confidence interval, Bootstrap Bias Correction and Acceleration HBCA confidence interval and ACNA confidence interval for the covariate-adjusted Youden index and its optimal cut-off point with/without normal error assumptions. Extensive simulation studies are conducted to evaluate the finite sample performance of various confidence intervals for the Youden index and its optimal cut-off point in the presence of covariates. To illustrate the application of our recommended methods, we apply the methods to a dataset on postprandial blood glucose measurements.

2.2 Linear Regression Models for the Test Results

2.2.1 The Youden Index and Its Associated Cut-off Point

Let X denote the non-diseased test result and Y denote the diseased test result. X and Y are linear functions of covariates based on:

$$X|\mathbf{Z} = \mathbf{z} = \boldsymbol{\beta}_1'\mathbf{z} + \varepsilon_1, \tag{2.1}$$

$$Y|\mathbf{Z} = \mathbf{z} = \boldsymbol{\beta}_2'\mathbf{z} + \varepsilon_2, \tag{2.2}$$

where $\mathbf{z} = (z_1, z_2, \cdots, z_p)'$ is a p-dimensional vector of covariates associated with the non-diseased and diseased test results, $\boldsymbol{\beta}_t = (\beta_{t1}, \beta_{t2}, \cdots, \beta_{tp})'$, $t = 1, 2$, are p-dimensional column vectors of unknown parameters, the error terms ε_i's are independent random variables, and $\varepsilon_t \sim N(0, \sigma_t^2)$ for $t = 1, 2$. At a given covariate $\mathbf{Z} = \mathbf{z}$, $X|\mathbf{Z} \sim N(\boldsymbol{\beta}_1'\mathbf{z}, \sigma_1^2)$, and $Y|\mathbf{Z} \sim N(\boldsymbol{\beta}_2'\mathbf{z}, \sigma_2^2)$. From equations (2.1) and (2.2), we can derive the covariate-adjusted YI at a given cut-off point c:

$$J(\mathbf{z}) = \max_c \{\Phi\left(\frac{\boldsymbol{\beta}_2'\mathbf{z} - c}{\sigma_2}\right) + \Phi\left(\frac{c - \boldsymbol{\beta}_1'\mathbf{z}}{\sigma_1}\right)\} - 1.$$

According to Schisterman and Perkins [29], the covariate-adjusted optimal cut-off point $c_0(\mathbf{z})$ is derived as

$$c_0(\mathbf{z}) = \frac{\boldsymbol{\beta}_1'\mathbf{z}(b^2 - 1) - a + b\sqrt{a^2 + (b^2 - 1)\sigma_1^2 \ln b^2}}{b^2 - 1}, \tag{2.3}$$

where $a = \beta_{\mathbf{2}}'\mathbf{z} - \beta_{\mathbf{1}}'\mathbf{z}, b = \frac{\sigma_2}{\sigma_1}$.

If $\sigma_1 = \sigma_2$, $c_0(\mathbf{z})$ can be replaced by the limit of (2.3) as $b \to 1$ which is

$$c_0(\mathbf{z}) = \frac{\boldsymbol{\beta}_1'\mathbf{z} + \boldsymbol{\beta}_2'\mathbf{z}}{2}. \tag{2.4}$$

Therefore, by substituting c with c_0, the covariate-adjusted YI is given by

$$J(\mathbf{z}) = \Phi\left(\frac{\boldsymbol{\beta}_2'\mathbf{z} - c_0(\mathbf{z})}{\sigma_2}\right) + \Phi\left(\frac{c_0(\mathbf{z}) - \boldsymbol{\beta}_1'\mathbf{z}}{\sigma_1}\right) - 1. \tag{2.5}$$

Suppose that $\{(\mathbf{z}_{i,x}', x_i) : i = 1, \cdots, m\}$ are random samples of "non-diseased" subjects from model (2.1) and $\{(\mathbf{z}_{j,y}', y_j) : j = 1, \cdots, n\}$ are random samples of "diseased" subjects from model (2.2). $\mathbf{z}_{i,x} = (z_{i1,x}, z_{i2,x}, \cdots, z_{ip,x})'$ and $\mathbf{z}_{j,y} = (z_{j1,y}, z_{j2,y}, \cdots, z_{jp,y})'$ are the corresponding covariates values in the "non-diseased" and "diseased" samples. We would like to estimate $J(\mathbf{z})$ and $c_0(\mathbf{z})$ at a given $\mathbf{z} = (z_1, z_2, \cdots, z_p)'$ based on these samples. Then, $\boldsymbol{\beta}_i$'s and σ_i^2's can be estimated by the following estimators based on the "non-diseased" and

"diseased" samples, respectively, i.e.,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_1 &= (\widetilde{\mathbf{Z}}_{\mathbf{x}}'\widetilde{\mathbf{Z}}_{\mathbf{x}})^{-1}\widetilde{\mathbf{Z}}_{\mathbf{x}}'\widetilde{\mathbf{X}}, \\
\widehat{\boldsymbol{\beta}}_2 &= (\widetilde{\mathbf{Z}}_{\mathbf{y}}'\widetilde{\mathbf{Z}}_{\mathbf{y}})^{-1}\widetilde{\mathbf{Z}}_{\mathbf{y}}'\widetilde{\mathbf{Y}}, \\
\widehat{\sigma}_1^2 &= (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} - \widehat{\boldsymbol{\beta}}_1\widetilde{\mathbf{Z}}_{\mathbf{x}}'\widetilde{\mathbf{X}})/(m-p), \\
\widehat{\sigma}_2^2 &= (\widetilde{\mathbf{Y}}'\widetilde{\mathbf{Y}} - \widehat{\boldsymbol{\beta}}_2\widetilde{\mathbf{Z}}_{\mathbf{y}}'\widetilde{\mathbf{Y}})/(n-p),
\end{aligned}$$

where $\widetilde{\mathbf{X}} = (x_1, \cdots, x_m)'$, $\widetilde{\mathbf{Y}} = (y_1, \cdots, y_n)'$ and

$$\widetilde{\mathbf{Z}}_{\mathbf{x}} = \begin{pmatrix} z_{11,x} & z_{12,x} & \cdots & z_{1p,x} \\ z_{21,x} & z_{22,x} & \cdots & z_{2p,x} \\ \vdots & \vdots & \vdots & \vdots \\ z_{m1,x} & z_{m2,x} & \cdots & z_{mp,x} \end{pmatrix}, \quad \widetilde{\mathbf{Z}}_{\mathbf{y}} = \begin{pmatrix} z_{11,y} & z_{12,y} & \cdots & z_{1p,y} \\ z_{21,y} & z_{22,y} & \cdots & z_{2p,y} \\ \vdots & \vdots & \vdots & \vdots \\ z_{n1,y} & z_{n2,y} & \cdots & z_{np,y} \end{pmatrix}.$$

Then by substituting above estimates for the corresponding unknown parameters in (2.3)-(2.5), we obtain the point estimators $\widehat{J}(\mathbf{z})$ and $\widehat{c}_0(\mathbf{z})$ for the covariate-adjusted YI along with its optimal cut-off point.

2.2.2 Generalized Confidence Intervals

Tsui and Weerahandi [19][25] introduced the generalized confidence interval. They provided the following definition for the generalized pivotal quantity (GPQ). Suppose that W is a random variable with a distribution depending on (θ, δ), where θ is a parameter of interest and δ is a nuisance parameter. Let w be the observed value of W. $R(W; w, \theta, \delta)$ is a function of W, w, θ, and δ. Then $R(W; w, \theta, \delta)$ is called a GPQ if it satisfies: 1. The distribution of $R(W; w, \theta, \delta)$ doesn't depend on the unknown parameters. 2. The value of $R(W; w, \theta, \delta)$ at $W = w$ is θ, which is the parameter of interest.

Note that $X|\mathbf{z}$ and $Y|\mathbf{z}$ are independently following normal distributions $N(\boldsymbol{\beta}_1'\mathbf{z}, \sigma_1^2)$ and $N(\boldsymbol{\beta}_2'\mathbf{z}, \sigma_2^2)$, respectively. Next, we will derive the corresponding GPQs for mean and variance functions: $\boldsymbol{\beta}_1'\mathbf{z}$, $\boldsymbol{\beta}_2'\mathbf{z}$, σ_1^2 and σ_2^2, at a given covariate $\mathbf{Z} = \mathbf{z}$.

We observe that $\boldsymbol{\beta}_t'\mathbf{z}$, t=1,2 can be consistently estimated by $\widehat{\boldsymbol{\beta}}_t'\mathbf{z}$, where $\hat{\boldsymbol{\beta}}_1'\mathbf{z}$ and $\hat{\boldsymbol{\beta}}_2'\mathbf{z}$ are linear combinations of $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ following multivariate normal distributions. $\widehat{\boldsymbol{\beta}}_t'\mathbf{z}$ is normally distributed following $\widehat{\boldsymbol{\beta}}_t'\mathbf{z} \sim N\left(\boldsymbol{\beta}_t'\mathbf{z}, \sigma_t^2 V_t\right)$ for $t = 1, 2$ with

$$Var(\widehat{\boldsymbol{\beta}}_t'\mathbf{z}) = \mathbf{z}'Var(\widehat{\boldsymbol{\beta}}_t)\mathbf{z} = \sigma_t^2\mathbf{z}'(\widetilde{\mathbf{Z}}_\mathbf{t}'\widetilde{\mathbf{Z}}_\mathbf{t})^{-1}\mathbf{z} \equiv \sigma_t^2 V_t. \tag{2.6}$$

Therefore, we can derive the GPQ for $\boldsymbol{\beta}_1'\mathbf{z}$ and $\boldsymbol{\beta}_2'\mathbf{z}$ as:

$$\begin{aligned} R_{\boldsymbol{\beta}_1'\mathbf{z}} &= \hat{\boldsymbol{\beta}}_1'\mathbf{z} - \frac{\hat{\boldsymbol{\beta}}_1'\mathbf{z} - \boldsymbol{\beta}_1'\mathbf{z}}{\sigma_1\sqrt{V_1}} \times \sigma_1\sqrt{V_1}\frac{e_x}{e_X} \\ &= \hat{\boldsymbol{\beta}}_1'\mathbf{z} - \frac{Z}{\sqrt{e_X^2/\sigma_1^2}} \times e_x\sqrt{V_1} \\ &= \hat{\boldsymbol{\beta}}_1'\mathbf{z} - T_{m-p}\sqrt{\frac{m}{m-p}} \times e_x\sqrt{V_1}, \end{aligned} \tag{2.7}$$

where $e_X = \left\{\frac{\sum_i (X_i - \bar{X})^2}{m}\right\}^{1/2}$ with $\bar{X} = \sum_i X_i/m$, e_x is the observed value of e_X, and T_{m-p} is a Student's t statistic with degrees of freedom $m - p$. And

$$R_{\boldsymbol{\beta}_2'\mathbf{z}} = \hat{\boldsymbol{\beta}}_2'\mathbf{z} - T_{n-p}\sqrt{\frac{n}{n-p}} \times e_y\sqrt{V_2}, \tag{2.8}$$

where $e_Y = \left\{\frac{\sum_j (Y_j - \bar{Y})^2}{n}\right\}^{1/2}$ with $\bar{Y} = \sum_j Y_j/n$, e_y is the observed value of e_Y, and T_{n-p} is a Student's t statistic with degrees of freedom $n - p$.

The GPQs for σ_1^2 and σ_2^2 are:

$$R_{\sigma_1^2} = \frac{\sigma_1^2}{e_X^2} \times e_x^2 = \frac{m e_x^2}{\chi^2_{m-p}}, \tag{2.9}$$

$$R_{\sigma_2^2} = \frac{\sigma_2^2}{e_Y^2} \times e_y^2 = \frac{n e_y^2}{\chi^2_{n-p}}, \tag{2.10}$$

respectively.

Then we obtain the GPQ for optimal cut-off point $c_0(\mathbf{z})$ by substituting $R_a, R_b, R_{\boldsymbol{\beta}_1'\mathbf{z}}$

and $R_{\sigma_1^2}$ for the corresponding quantities $a, b, \boldsymbol{\beta}_1'\mathbf{z}$ and σ_1^2 in (2.3) and (2.4):

$$R_{c_0} = \frac{R_{\boldsymbol{\beta}_1'\mathbf{z}}(R_b^2 - 1) - R_a + R_b\sqrt{R_a^2 + (R_b^2 - 1)R_{\sigma_1^2}\ln R_b^2}}{R_b^2 - 1}. \tag{2.11}$$

where R_a and R_b are the GPQs for a and b, respectively,

$R_a = R_{\boldsymbol{\beta}_2'\mathbf{z}} - R_{\boldsymbol{\beta}_1'\mathbf{z}}$, $R_b = \frac{R_{\sigma_2}}{R_{\sigma_1}}$, and $R_{\sigma_t} = \sqrt{R_{\sigma_t^2}}$ for $t = 1, 2$.

When $\sigma_1^2 = \sigma_2^2$,

$$R_{c_0} = \frac{R_{\boldsymbol{\beta}_1'\mathbf{z}} + R_{\boldsymbol{\beta}_2'\mathbf{z}}}{2}. \tag{2.12}$$

Therefore, the GPQ for Youden index $J(\mathbf{z})$ is

$$R_J = \Phi\left(\frac{R_{\boldsymbol{\beta}_2'\mathbf{z}} - R_{c_0}}{R_{\sigma_2}}\right) + \Phi\left(\frac{R_{c_0} - R_{\boldsymbol{\beta}_1'\mathbf{z}}}{R_{\sigma_1}}\right). \tag{2.13}$$

In order to construct $100(1-\alpha)\%$ level generalized confidence intervals for the covariate-adjusted YI along with its optimal cut-off point, we apply the following algorithm procedure(see also Tian [22], Lai *et al.*[24]):

1. Compute $e_x = \left\{\frac{\sum_i (x_i - \bar{x})^2}{m}\right\}^{1/2}$, $e_y = \left\{\frac{\sum_j (y_j - \bar{y})^2}{n}\right\}^{1/2}$, $\hat{\boldsymbol{\beta}}_1'\mathbf{z}$ and $\hat{\boldsymbol{\beta}}_2'\mathbf{z}$.

2. Let k be the loop number, we choose $K = 1000$ as the total number of iterations. Then, for each loop, $k = 1, \ldots, K$,

   - Generate T_{m-p} and T_{n-p} from Student's t-distribution with degrees of freedom $m - p$ and $n - p$, respectively;
   - Generate χ^2_{m-p} and χ^2_{n-p} from χ^2 distribution with degrees of freedom $m - p$ and $n - p$, respectively;
   - Compute $R_{\boldsymbol{\beta}_1'\mathbf{z}}, R_{\boldsymbol{\beta}_2'\mathbf{z}}, R_{\sigma_1}$, and R_{σ_2} according to equations (2.7)-(2.10);
   - Compute $R_{c_0,k}$ following (2.11) or (2.12);
   - Compute $R_{J,k}$ following (2.13).

(end k loop)

3. Compute the $100\alpha/2$-th percentile $R_{J,\alpha/2}$ and the $100(1-\alpha/2)$-th percentile $R_{J,(1-\alpha/2)}$ of $\{R_{J,1}, R_{J,2}, \ldots, R_{J,K}\}$. Then, $\left(R_{J,\alpha/2}, R_{J,(1-\alpha/2)}\right)$ is a $100(1-\alpha)\%$ level confidence interval for the covariate-adjusted YI.

4. Compute the $100\alpha/2$-th percentile $R_{c_0,\alpha/2}$ and the $100(1-\alpha/2)$-th percentile $R_{c_0,(1-\alpha/2)}$ of $\{R_{c_0,1}, R_{c_0,2}, \ldots, R_{c_0,K}\}$. Then, $\left(R_{c_0,\alpha/2}, R_{c_0,(1-\alpha/2)}\right)$ is a $100(1-\alpha)\%$ level confidence interval for the covariate-adjusted optimal cut-off point.

The normality assumption on the test results $X|\mathbf{z}$ and $Y|\mathbf{z}$ may not be necessarily satisfied in many applications. For test results with non-normal distributions, the transformed test results could follow normal distributions after using the following Box-Cox transformation:

$$X^{(\lambda)}|\mathbf{z} = \begin{cases} \frac{X^{\lambda}|\mathbf{z}-1}{\lambda}, & \lambda \neq 0 \\ log(X^{\lambda}|\mathbf{z}), & \lambda = 0 \end{cases}, \qquad Y^{(\lambda)}|\mathbf{z} = \begin{cases} \frac{Y^{\lambda}|\mathbf{z}-1}{\lambda}, & \lambda \neq 0 \\ log(Y|\mathbf{z}), & \lambda = 0 \end{cases},$$

where the power constant λ can be obtained by maximizing the likelihood functions of the transformed test results. Then, the proposed GPQ method can be applied to the transformed test results for inferences on the YI and its associated optimal cutoff.

2.3 Heteroscedastic Regression Models for the Test Results

In section 2.2, we employ linear regressions with normal errors to model the covariate effects on test results. For test results with non-normal or unknown distributions, Yao et al.[10] motivated the use of non-parametric heteroscedastic regression models for test results. Here we utilize the same models as in Yao, et al. [10], and assume that

$$X|\mathbf{Z} = \mathbf{z} = \mu_1(\mathbf{z}) + \sqrt{\nu_1(\mathbf{z})}\epsilon_1, \tag{2.14}$$

$$Y|\mathbf{Z} = \mathbf{z} = \mu_2(\mathbf{z}) + \sqrt{\nu_2(\mathbf{z})}\epsilon_2, \tag{2.15}$$

where $\mathbf{z}$ represents the vector of covariates, ϵ_1 and ϵ_2 are independent standard errors having mean zero and standard deviation one, the range of the variance functions $\nu_1(\mathbf{z})$ and $\nu_2(\mathbf{z})$ is restricted in $\Re^+$ and finite for all $\mathbf{z} \in \Re^p$. In addition, let $F_{\mathbf{Z}}$ and $G_{\mathbf{Z}}$ denote the cumulative distribution functions (c.d.f.) of X and Y at given $\mathbf{Z}$ respectively, $F^*(\cdot)$ and $G^*(\cdot)$ denote the c.d.f. of ϵ_1 and ϵ_2 respectively. Here, the error distributions F^* and G^* are assumed to be independent of $\mathbf{Z}$. We further assume that for any given covariate value $\mathbf{Z} = \mathbf{z}$, $P(Y > X|\mathbf{Z} = \mathbf{z}) \geq 0.5$, which is equivalent to $\mu_1(\mathbf{z}) < \mu_2(\mathbf{z})$ if F^* and G^* are symmetric distributions about 0. This assumption ensures that the value of the YI with given covariate information is between 0 and 1 inclusive.

2.3.1 Covariate-adjusted Youden Index with the Normal Error Assumption

With the covariate $\mathbf{Z}$, both YI and its associated optimal cut-off point are dependent on $\mathbf{Z}$. The YI at given $\mathbf{Z} = \mathbf{z}$ is

$$\begin{aligned} J(\mathbf{z}) &= \max_c \{P(Y \geq c|\mathbf{Z} = \mathbf{z})) + P(X \leq c|\mathbf{Z} = \mathbf{z}) - 1\} && (2.16) \\ &= P(X \leq c_o(\mathbf{z})|\mathbf{Z} = \mathbf{z}) - P(Y \leq c_o(\mathbf{z})|\mathbf{Z} = \mathbf{z}) && (2.17) \\ &= F_{\mathbf{Z}}(c_o(\mathbf{z})) - G_{\mathbf{Z}}(c_o(\mathbf{z})), && (2.18) \end{aligned}$$

where $c_o(\mathbf{z})$ is the optimal cut-off point at given $\mathbf{z}$.

If the errors ϵ_1 and ϵ_2 are assumed to be normally distributed in models (2.14) and (2.15), the YI at $\mathbf{Z} = \mathbf{z}$ can be expressed as

$$J_N(\mathbf{z}) = \Phi\left(\frac{\mu_2(\mathbf{z}) - c_o(\mathbf{z})}{\sqrt{\nu_2(\mathbf{z})}}\right) - \Phi\left(\frac{\mu_1(\mathbf{z}) - c_o(\mathbf{z})}{\sqrt{\nu_1(\mathbf{z})}}\right) \equiv \theta_{2N} - \theta_{1N}, \quad (2.19)$$

where $J_N(\mathbf{z})$ stands for $J(\mathbf{z})$ under the normal distributional assumption for the errors. With the assumption that $\mu_2(\mathbf{z}) > \mu_1(\mathbf{z})$, $c_o(\mathbf{z})$ has the following closed form:

$$c_o(\mathbf{z}) = \frac{\mu_1(\mathbf{z})(b^2 - 1) - a + b\sqrt{a^2 + (b^2 - 1)\nu_1(\mathbf{z})ln(b^2)}}{(b^2 - 1)}, \quad (2.20)$$

where $a = \mu_2(\mathbf{z}) - \mu_1(\mathbf{z})$, $b = \sqrt{\nu_2(\mathbf{z})}/\sqrt{\nu_1(\mathbf{z})}$. When $b = 1$, we have

$$c_o(\mathbf{z}) = \frac{\mu_1(\mathbf{z}) + \mu_2(\mathbf{z})}{2}. \tag{2.21}$$

Under models (2.14)-(2.15), the mean and variance functions μ_1, μ_2, ν_1, and ν_2 can be consistently estimated via the local linear or kernel regression techniques. Let $\widehat{\mu}_1$, $\widehat{\mu}_2$, $\widehat{\nu}_1$, and $\widehat{\nu}_2$ be the local linear estimates for μ_1, μ_2, ν_1, and ν_2 (see Fan and Gijbels, 1996) respectively, and $c_{oN}(\mathbf{z})$ be the plug-in estimate of $c_o(\mathbf{z})$. In practice, Hengartner, Wegkamp and Matzner-Løber's [30] method can be used to select bandwidths for these local linear regression estimates. Then the estimator for the covariate-adjusted YI can be defined as follows:

$$\widehat{J}_N(\mathbf{z}) = \Phi\left(\frac{\mu_2(\mathbf{z}) - \widehat{c}_{oN}(\mathbf{z})}{\sqrt{\widehat{\nu}_2(\mathbf{z})}}\right) - \Phi\left(\frac{\widehat{\mu}_1(\mathbf{z}) - \widehat{c}_{oN}(\mathbf{z})}{\sqrt{\widehat{\nu}_1(\mathbf{z})}}\right) \equiv \widehat{\theta}_{2N} - \widehat{\theta}_{1N}. \tag{2.22}$$

2.3.2 Covariate-adjusted Youden Index without the Normal Error Assumption

In this section, we assume that the error distributions of ϵ_t's $(t = 1, 2)$ in models (2.14)-(2.15) are unknown. As mentioned in the Section 2.2, $x_i, i = 1 \cdots m$ and $y_j, j = 1 \cdots n$ are test results of random samples from "non-diseased" and "diseased" subjects respectively. $z_{i,x}$ and $z_{j,y}$ are given covariates based on these samples. We want to estimate $J(\mathbf{z})$ at given $\mathbf{z}$.

To estimate $J(\mathbf{z})$ at given $\mathbf{z}$, we have to estimate the mean functions $\mu_i(\mathbf{z})$'s, the variance functions $\nu_i(\mathbf{z})$'s and the error distributions. We can easily estimate the mean and variance functions by using the local linear or kernel regression methods. However it is difficult to obtain good estimates for the error distributions in the heteroscedastic regression models. Instead of using a complex distribution estimation (e.g., a kernel distribution estimation), we employ the following procedure to estimate $J(\mathbf{z})$ (see also Yao, et al.[10]).

1. Find the local linear estimates $\widehat{\mu}_1$, $\widehat{\mu}_2$, $\widehat{\nu}_1$, and $\widehat{\nu}_2$ for μ_1, μ_2, ν_1, and ν_2 (see Fan and Gijbels, [31]).

2. Find the standardized residuals:

$$\widehat{\epsilon}_{i,1} = \frac{x_i - \widehat{\mu}_1(\mathbf{z}_{i,x})}{\sqrt{\widehat{\nu}_1(\mathbf{z}_{i,x})}}, \quad \widehat{\epsilon}_{j,2} = \frac{y_j - \widehat{\mu}_2(\mathbf{z}_{j,y})}{\sqrt{\widehat{\nu}_2(\mathbf{z}_{j,y})}}.$$

3. Estimate test values at given $\mathbf{Z} = \mathbf{z}$ as follows:

$$\widehat{x}_{i,\mathbf{z}} = \widehat{\mu}_1(\mathbf{z}) + \sqrt{\widehat{\nu}_1(\mathbf{z})}\widehat{\epsilon}_{i,1}, \quad \widehat{y}_{j,\mathbf{z}} = \widehat{\mu}_2(\mathbf{z}) + \sqrt{\widehat{\nu}_2(\mathbf{z})}\widehat{\epsilon}_{j,2}.$$

Then, the covariate-adjusted YI $J(\mathbf{z})$ can be estimated by

$$\begin{aligned}
\widehat{J}_E(\mathbf{z}) &= \max_c \left[n^{-1} \sum_{j=1}^{n} I(\widehat{y}_{j,\mathbf{z}} \geq c) - m^{-1} \sum_{i=1}^{m} I(\widehat{x}_{i,\mathbf{z}} \geq c) \right] \\
&= n^{-1} \sum_{j=1}^{n} I(\widehat{y}_{j,\mathbf{z}} \geq \widehat{c}_{oE}(\mathbf{z})) - m^{-1} \sum_{i=1}^{m} I(\widehat{x}_{i,\mathbf{z}} \geq \widehat{c}_{oE}(\mathbf{z})) \\
&\equiv \widehat{\theta}_{2E} - \widehat{\theta}_{1E} \equiv \widehat{\theta},
\end{aligned}$$

where $I(\cdot)$ is the indicator function, and $\widehat{c}_{oE}(\mathbf{z})$ is an empirical estimator for the optimal cut-off point and defined as $\widehat{c}_{oE}(\mathbf{z}) =$ median of $\hat{C}(\mathbf{z})$ with

$$\hat{C}(\mathbf{z}) = \left\{ c : \max_c \left[n^{-1} \sum_{j=1}^{n} I(\widehat{y}_{j,\mathbf{z}} \geq c) - m^{-1} \sum_{i=1}^{m} I(\widehat{x}_{i,\mathbf{z}} \geq c) \right] \right\}.$$

2.3.3 Confidence Intervals for the Covariate-adjusted Youden Index

Zhou and Qin [11] studied the asymptotic properties of the estimator $\widehat{J}(\mathbf{z})$ for the covariate-adjusted YI under heteroscedastic regression models with/without the normal error assumption. Here, we focus on construction of confidence intervals for the covariate-adjusted YI.

From (3.3), since

$$\begin{aligned} J(\mathbf{z}) &= P(Y \geq c_o(\mathbf{z})|\mathbf{Z}=\mathbf{z}) + P(X \leq c_o(\mathbf{z})|\mathbf{Z}=\mathbf{z}) - 1 && (2.23)\\ &= P(Y \geq c_o(\mathbf{z})|\mathbf{Z}=\mathbf{z}) - P(X \geq c_o(\mathbf{z})|\mathbf{Z}=\mathbf{z}) && (2.24)\\ &\equiv \theta_2 - \theta_1 \equiv \theta, && (2.25) \end{aligned}$$

we can see that the covariate-adjusted YI $J(\mathbf{z})$ is the difference between two unknown proportions θ_2 and θ_1, where $\theta_1 \equiv P(X \geq c_o(\mathbf{z})|\mathbf{Z}=\mathbf{z})$, $\theta_2 \equiv P(Y \geq c_o(\mathbf{z})|\mathbf{Z}=\mathbf{z})$. Since $c_o(\mathbf{z})$ can be consistently estimated by $\widehat{c}_o(\mathbf{z}) = \widehat{c}_{oN}(\mathbf{z})$ or $\widehat{c}_{oE}(\mathbf{z})$, θ_1 can be consistently estimated by $\widehat{\theta}_1 = \widehat{\theta}_{1N}$ or $\widehat{\theta}_{1E}$, and θ_2 can be consistently estimated by $\widehat{\theta}_2 = \widehat{\theta}_{2N}$ or $\widehat{\theta}_{2E}$, respectively. Hence, $J(\mathbf{z})$ can be consistently estimated by $\widehat{J}(\mathbf{z}) = \widehat{J}_N(\mathbf{z})$ or $\widehat{J}_E(\mathbf{z})$ with/without the normal error assumption, respectively.

Under the assumption that the test results from the non-diseased group are independent of the test results from the diseased group, the variance of $\widehat{J}(\mathbf{z})$ can be consistently estimated by

$$\widehat{Var}(J(\mathbf{z})) = \widehat{Var}(\hat{\theta}_2 - \hat{\theta}_1) = \widehat{Var}(\hat{\theta}_2) + \widehat{Var}(\hat{\theta}_1),$$

where $\widehat{Var}(\hat{\theta}_1) = \hat{\theta}_1(1-\hat{\theta}_1)/n$ and $\widehat{Var}(\hat{\theta}_2) = \hat{\theta}_2(1-\hat{\theta}_2)/m$ are consistent estimates for the variance of $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively. Therefore, a $(1-\alpha)$-th Wald confidence interval for the covariate-adjusted YI can be constructed as follows:

$$\left(J(\mathbf{z}) - z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\theta}_1) + \widehat{Var}(\hat{\theta}_2)},\ J(\mathbf{z}) + z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\theta}_1) + \widehat{Var}(\hat{\theta}_2)} \right), \qquad (2.26)$$

where $z_{\alpha/2}$ is the upper $\frac{\alpha}{2}$-th quantile of the standard normal distribution.

Our simulation studies show that this Wald confidence interval has poor small sample performance. In order to improve the performance of the Wald confidence interval, we use the MOVER method (See Zou, [32]) to construct new hybrid confidence intervals for the covariate-adjusted YI and its optimal cut-off point.

Let l_t and u_t $(t = 1, 2)$ be the lower and upper limits of a given $100(1-\alpha)\%$ two-sided confidence interval for θ_t, respectively. Then,

$$l_t = \hat{\theta}_t - z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\theta}_t)}, \quad u_t = \hat{\theta}_t + z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\theta}_t)},$$

which implies that the variance of $\hat{\theta}_t$ can be estimated by $\widehat{Var}_l(\hat{\theta}_t) = (\hat{\theta}_t - l_t)^2/z^2_{\alpha/2}$ and $\widehat{Var}_u(\hat{\theta}_t) = (u_t - \hat{\theta}_t)^2/z^2_{\alpha/2}$. After plugging these variance estimates back to equation (2.26), we get a hybrid confidence interval for the covariate-adjusted YI:

$$\left(J(\mathbf{z}) - \sqrt{(\hat{\theta}_2 - l_2)^2 + (u_1 - \hat{\theta}_1)^2},\ J(\mathbf{z}) + \sqrt{(u_2 - \hat{\theta}_2)^2 + (\hat{\theta}_1 - l_1)^2}\right).$$

Here, we propose methods (i) and (ii) to construct the two-sided confidence interval (l_t, u_t) for θ_t.

(i) *The Wilson Score Method*

$$l_1 = \frac{\hat{\theta}_1 + \frac{z^2_{\alpha/2}}{2m} - z_{\alpha/2}\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{m} + \frac{z^2_{\alpha/2}}{4m^2}}}{1 + z^2_{\alpha/2}/m}, \quad u_1 = \frac{\hat{\theta}_1 + \frac{z^2_{\alpha/2}}{2m} + z_{\alpha/2}\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{m} + \frac{z^2_{\alpha/2}}{4m^2}}}{1 + z^2_{\alpha/2}/m},$$

$$l_2 = \frac{\hat{\theta}_2 + \frac{z^2_{\alpha/2}}{2n} - z_{\alpha/2}\sqrt{\frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + z^2_{\alpha/2}/n}, \quad u_2 = \frac{\hat{\theta}_2 + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + z^2_{\alpha/2}/n}.$$

The hybrid confidence intervals based on this method are called *Hybrid Wilson Score* (HWS) intervals for the covariate-adjusted YI. We use HWS-N or HWS-E denote the corresponding HWS interval for $J(\mathbf{z})$ when $\theta_t = \theta_{tN}$ or θ_{tE} for $t = 1, 2$.

(ii) *The Agresti-Coull Method*

$$l_1 = \tilde{\theta}_1 - z_{\alpha/2}\sqrt{\frac{\tilde{\theta}_1(1-\tilde{\theta}_1)}{m + z^2_{\alpha/2}}}, \qquad u_1 = \tilde{\theta}_1 + z_{\alpha/2}\sqrt{\frac{\tilde{\theta}_1(1-\tilde{\theta}_1)}{m + z^2_{\alpha/2}}},$$

where $\tilde{\theta}_1 = \tilde{\theta}_{1N} \equiv \frac{\hat{\theta}_{1N} + z^2_{\alpha/2}/(2m)}{1 + z^2_{\alpha/2}/m}$, or $\tilde{\theta}_{1E} \equiv \frac{\sum_{i=1}^m I(\hat{x}_{i,\mathbf{z}} \geq \hat{c}_{oE}(\mathbf{z})) + z^2_{\alpha/2}/2}{m + z^2_{\alpha/2}}$,

$$l_2 = \tilde{\theta}_2 - z_{\alpha/2}\sqrt{\frac{\tilde{\theta}_2(1-\tilde{\theta}_2)}{n+z^2_{\alpha/2}}}, \qquad u_2 = \tilde{\theta}_2 + z_{\alpha/2}\sqrt{\frac{\tilde{\theta}_2(1-\tilde{\theta}_2)}{n+z^2_{\alpha/2}}},$$

where $\tilde{\theta}_2 = \tilde{\theta}_{2N} \equiv \frac{\widehat{\theta}_{2N}+z^2_{\alpha/2}/(2n)}{1+z^2_{\alpha/2}/n}$, or $\tilde{\theta}_{2E} \equiv \frac{\sum_{j=1}^{n} I(\widehat{y}_{j,\mathbf{z}} \geq \widehat{c}_{oE}(\mathbf{z}))+z^2_{\alpha/2}/2}{n+z^2_{\alpha/2}}$.

The hybrid confidence intervals for $J(\mathbf{z})$ based on this method are called *Hybrid Agresti-Coull* (HAC) intervals for the covariate-adjusted YI. We use HAC-N or HAC-E denote the corresponding HAC interval for $J(\mathbf{z})$ with/without the normal error assumption.

Another confidence interval for $\theta = J(\mathbf{z})$, or $c_o(\mathbf{z})$ is the bootstrap *Bias Correction and Acceleration* (BCa) interval defined as

$$(\theta^{*([B\gamma_1])}, \theta^{*([B\gamma_2])}),$$

where $\{\theta^{*b} : b = 1, 2, \cdots, B\}$ are B ($B \geq 500$ is recommended) bootstrap copies of $\widehat{\theta} = \widehat{J}(\mathbf{z})$, or $c_o(\mathbf{z})$, $\theta^{*([B\gamma_1])}$ and $\theta^{*([B\gamma_2])}$ are the $[B\gamma_1]$-th and $[B\gamma_2]$-th ordered values of $\{\theta^{*b}\}$'s respectively, and

$$\gamma_1 = \Phi\left(q + \frac{q+z_{\alpha/2}}{1-p(q+z_{\alpha/2})}\right), \quad \gamma_2 = \Phi\left(q + \frac{q+z_{1-\alpha/2}}{1-p(q+z_{1-\alpha/2})}\right),$$

with $p = \frac{1}{6}\sum_{j=1}^{m+n}\phi_j^3/(\sum_{j=1}^{m+n}\phi_j^2)^{\frac{3}{2}}$, $q = \Phi^{-1}(\frac{1}{B}\sum_{b=1}^{B} I(\theta_b^* \leq \hat{\theta}))$, $\phi_j = \hat{\theta}_{(\cdot)} - \hat{\theta}_{(-j)}$, $\hat{\theta}_{(-j)}$ being $\hat{\theta}$ computed by deleting the j-th observation in the combined sample from the non-diseased and diseased populations, and $\hat{\theta}_{(\cdot)} = \frac{1}{m+n}\sum_{j=1}^{m+n}\hat{\theta}_{(-j)}$.

Under the *Heteroscedastic Regression Models*, the confidence intervals for $J(\mathbf{z})$ and $c_o(\mathbf{z})$ based on this BCa method are called the HBCA intervals for the covariate-adjusted YI along with its optimal cut-off point. We use HBCA-N or HBCA-E denote the corresponding HBCA intervals for $J(\mathbf{z})$ and $c_o(\mathbf{z})$ with/without the normal error assumption.

Zhou and Qin (2015) also proposed a bootstrap-based interval for $J(\mathbf{z})$. For comparison

with our proposed intervals (see simulation studies in next section), we summarize their bootstrap procedure here.

Let

$$\widehat{J}_{AC}(\mathbf{z}) = \frac{\sum_{j=1}^{n} I(\widehat{y}_{j,\mathbf{z}} \geq \widehat{c}_{oE}(\mathbf{z})) + z_{\alpha/2}^2/2}{n + z_{\alpha/2}^2} - \frac{\sum_{i=1}^{m} I(\widehat{x}_{i,\mathbf{z}} \geq \widehat{c}_{oE}(\mathbf{z})) + z_{\alpha/2}^2/2}{m + z_{\alpha/2}^2}$$

where $\widehat{c}_{oE}(\mathbf{z})$ = median of $\hat{C}(\mathbf{z})$ is defined in section 2.2.

(1). Find a bootstrap sample of size m, $\widehat{x}^*_{i,\mathbf{z}}$'s, with replacement from $\widehat{x}_{i,\mathbf{z}}$'s, and a bootstrap sample of size n, $\widehat{y}^*_{j,\mathbf{z}}$'s, with replacement from $\widehat{y}_{j,\mathbf{z}}$'s.

(2). Calculate the bootstrap version of $\widehat{J}_{AC}(\mathbf{z})$

$$\widehat{J}^*_{AC}(\mathbf{z}) = \frac{\sum_{j=1}^{n} I(\widehat{y}^*_{j,\mathbf{z}} \geq \widehat{c}^*_{oE}(\mathbf{z})) + z_{\alpha/2}^2/2}{n + z_{\alpha/2}^2} - \frac{\sum_{i=1}^{m} I(\widehat{x}^*_{i,\mathbf{z}} \geq \widehat{c}^*_{oE}(\mathbf{z})) + z_{\alpha/2}^2/2}{m + z_{\alpha/2}^2},$$

where $\widehat{c}^*_{oE}(\mathbf{z})$ is the bootstrap version of $\widehat{c}_{oE}(\mathbf{z})$.

(3). Repeat step (1) and step (2) B ($B \geq 500$ is recommended) times to obtain the set of bootstrap replications $\{\widehat{J}^{*b}_{AC}(\mathbf{z}) : b = 1, 2, \ldots, B\}$ and $\{\widehat{c}^{*b}_{oE}(\mathbf{z}) : b = 1, 2, \ldots, B\}$.

Then, the bootstrap variance estimators $V^*(\widehat{J}_{AC}(\mathbf{z}))$ and $V^*(\widehat{c}_{oE}(\mathbf{z}))$ are defined as

$$V^*(\widehat{J}_{AC}(\mathbf{z})) = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{J}^{*b}_{AC}(\mathbf{z}) - \bar{J}^*_{AC}(\mathbf{z}))^2$$

$$V^*(\widehat{c}_{oE}(\mathbf{z})) = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{c}^{*b}_{oE}(\mathbf{z}) - \bar{c}^*_{oE}(\mathbf{z}))^2$$

where $\bar{J}^*_{AC}(\mathbf{z}) = \frac{1}{B}\sum_{b=1}^{B}\widehat{J}^{*b}_{AC}(\mathbf{z})$, and $\bar{c}^*_{oE}(\mathbf{z}) = \frac{1}{B}\sum_{b=1}^{B}\widehat{c}^{*b}_{oE}(\mathbf{z})$.

The bootstrap-based intervals (hereafter ACNA interval) for $J(\mathbf{z})$ and $c_o(\mathbf{z})$ are defined as

$$\left(\widehat{J}_{AC}(\mathbf{z}) - z_{\alpha/2}\sqrt{V^*(\widehat{J}_{AC}(\mathbf{z}))},\ \widehat{J}_{AC}(\mathbf{z}) + z_{\alpha/2}\sqrt{V^*(\widehat{J}_{AC}(\mathbf{z}))}\right), \tag{2.27}$$

$$\left(\widehat{c}_{oE}(\mathbf{z}) - z_{\alpha/2}\sqrt{V^*(\widehat{c}_{oE}(\mathbf{z}))},\ \widehat{c}_{oE}(\mathbf{z}) + z_{\alpha/2}\sqrt{V^*(\widehat{c}_{oE}(\mathbf{z}))}\right), \tag{2.28}$$

respectively.

2.4 Simulation Studies

In this section, we conduct two simulation studies to evaluate the finite sample performance of the confidence intervals proposed in sections 2.2 - 2.3.

In the first simulation study, we compare the proposed generalized confidence interval (GPQ interval) with the BCa interval and two bootstrap-based intervals for the covariate-adjusted YI along with the optimal cut-off point $c_0(\mathbf{z})$ in terms of interval lengths and coverage probabilities under the linear regression models with normal errors for test results. The computation procedure of the bootstrap-based intervals is summarized as follows:

1. Draw a bootstrap resample $\{(\mathbf{z}'^{*}_{i,x}, x^{*}_{i}) : i = 1, \cdots, m\}$ from the "non-diseased" sample $\{(\mathbf{z}'_{i,x}, x_i) : i = 1, \cdots, m\}$, and a bootstrap resample $\{(\mathbf{z}'^{*}_{j,y}, y^{*}_{j}) : j = 1, \cdots, n\}$ from the "diseased" sample $\{(\mathbf{z}'_{j,y}, y_j) : j = 1, \cdots, n\}$, respectively.

2. For $\hat{\theta} = \widehat{J}(\mathbf{z})$ and $\hat{c}_0(\mathbf{z})$, compute the bootstrap copy θ^* of $\hat{\theta}$ from (2.5) and (2.3), respectively.

3. Repeat the first two steps B times to obtain the bootstrap replications $\{\theta^{*b} : b = 1, 2, \cdots, B\}$. Then, the bootstrap estimator $V^*(\hat{\theta})$ for the variance of $\hat{\theta}$ is defined as

$$V^*(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\theta^{*b} - \overline{\theta}^*)^2,$$

   where $\overline{\theta}^* = B^{-1}\sum_{b=1}^{B}\theta^{*b}$.

Two $(1-\alpha)100\%$ $(0 < \alpha < 1)$ level bootstrap-based intervals for θ $(= J(\mathbf{z})$, or $c_0(\mathbf{z}))$ can be constructed as follows:

The first one, called BTI interval, is a bootstrap percentile interval for θ defined by

$$(\theta^{*([B\alpha/2])}, \theta^{*([B(1-\alpha/2)])}),$$

where $\theta^{*([B\alpha/2])}$ and $\theta^{*([B(1-\alpha/2)])}$ are the $\alpha/2$-th and $(1-\alpha/2)$-th quantiles of $\{\theta^{*b} : b = 1, 2, \cdots, B\}$, respectively.

The second one, called BTII interval, for θ is defined as

$$(\widehat{\theta} - z_{\alpha/2}\sqrt{V^*(\hat{\theta})}, \widehat{\theta} + z_{\alpha/2}\sqrt{V^*(\hat{\theta})}),$$

where $\hat{\theta}$ $(= \widehat{J}(\mathbf{z})$, or $\hat{c}_0(\mathbf{z}))$ is defined in section 2.1.

In the simulation study, the "non-diseased" sample $\{(\mathbf{z}'_{i,x}, x_i) : i = 1, \cdots, m\}$ and the "diseased" sample $\{(\mathbf{z}'_{j,y}, y_j) : j = 1, \cdots, n\}$ are generated from the following linear regression models, respectively:

Model 1: Linear Regression Models

$$X|Z = 6 + 1.5Z + \varepsilon_1, \tag{2.29}$$

$$Y|Z = 7.2 + 2.4Z + \varepsilon_2, \tag{2.30}$$

where Z follows the uniform distribution on $[1, 5]$, ε_t's follow $N(0, \sigma_i^2)$ $(t = 1, 2)$ with $\sigma_1 = 2$ and $\sigma_2 = 1.5$, and ε_1 is independent of ε_2. We choose sample sizes $(m, n) = (50, 50)$, $(100, 100)$, and $(30, 50)$, respectively. Using the simulated samples, we calculate 95% level confident intervals for the covariate-adjusted YI and the cut-off point $c_0(z)$ at a given z where z's are chosen to be 40 evenly distributed points in $[1, 5]$. For computation of the average upper and lower bounds of these confident intervals, we generate $N = 1000$ "non-diseased" and "diseased" samples, respectively. We choose $K = 1000$ for the calculation of the GPQ-based intervals, and $B = 1000$ for the calculation of BCa, BTI and BTII intervals. Figure 2.1 and 2.2 display the 95% level pointwise confidence bands and coverage probabilities of the confidence intervals for the covariate-adjusted YI and the optimal cut-off point respectively using BCa, GPQ, BTI and BTII methods under **Model 1**.

From Figure 2.1, we observe that when sample sizes get bigger, the coverage probabilities of all intervals for the covariate-adjusted YI are closer to the 95% nominal level, and the average lengths of the intervals become shorter. Among the three bootstrap (BCa,

BTI, BTII) intervals, BCa intervals have the coverage probabilities farther from the 95% nominal level than BTI and BTII intervals. When sample sizes are small, the coverage probabilities of BCa intervals can be far below the 95% nominal level at some values of the covariate. Comparing the GPQ method with the bootstrap-based methods, we can see that the GPQ-based intervals have coverage probabilities closer to the 95% nominal level than the bootstrap-based intervals, and the GPQ method is stable for all cases considered here. All the confidence intervals have comparable average lengths. In all cases, the computation time for the bootstrap-based intervals is far longer than that for the GPQ-based intervals. As for intervals of the optimal cut-off point, similar conclusions can be reached by observing the results in Figure 2.2. Hence we recommend the GPQ method for inferences on the covariate-adjusted YI and the associated optimal cut-off point under the linear regression models with normal errors for test results. In the second simulation study, we examine the finite sample performances of the HWS (HWS-N, HWS-E) intervals, HAC (HAC-N, HAC-E) intervals, HBCA (HBCA-N, HBCA-E) intervals, and ACNA intervals for the covariate-adjusted YI at given $Z = z$ under the following heteroscedastic regression models. At the same time, we examine the finite sample performances of the HBCA (HBCA-N, HBCA-E) intervals and ACNA intervals for the optimal cut-off point $c_o(z)$ under the same regression models.

Model 2: Heteroscedastic Regression Models

$$\begin{aligned} X|Z &= 6 + 1.5Z + 1.5\sin(Z) + \sqrt{0.4 + \Phi(2Z-6)}\epsilon_1, \\ Y|Z &= 7.2 + 1.5Z + 1.5\sin(Z) + \sqrt{Z - 0.8} + \sqrt{1.2 + \Phi(2Z-6)}\epsilon_2, \end{aligned}$$

where Z follows the uniform distribution on $[1, 5]$, both ϵ_1 and ϵ_2 are random error with mean zero and variance one, and Φ is the c.d.f. of the standard normal distribution. We evaluate these intervals under the scenario with/without normality assumption for the error distributions. In the first scenario, both ϵ_1 and ϵ_2 follow the standard normal distribution. In the second scenario, ϵ_1 and ϵ_2 follow a scaled student t-distribution $t_4/\sqrt{2}$ with degree of freedom 4.

Similar to the first simulation study, $N = 1000$ "non-diseased" and "diseased" samples are generated from **Model 2** with/without the normality assumption for the error distributions, respectively. We choose the sample sizes $(m, n) = (50, 50)$, $(100, 100)$, and $(80, 100)$, and $B = 1000$ for calculating 95% level pointwise confidence bands and coverage probabilities of the intervals for the covariate-adjusted YI and the optimal cut-off point $c_o(z)$ at a given z where z's are chosen to be 40 evenly distributed points in $[1, 5]$. Figures 2.3 - 2.6 display the 95% level pointwise confidence bands for the covariate-adjusted YI and the optimal cut-off point, and the corresponding coverage probabilities of the HWS (HWS-N, HWS-E) intervals, HAC (HAC-N, HAC-E) intervals, HBCA (HBCA-N, HBCA-E) intervals, and ACNA intervals.

When the errors ϵ_t's distributions are the standard normal distribution, from Figure 2.3 - 2.4, we can see that the coverage probabilities of HBCA-N intervals are much more closer to the 95% nominal level even when z is near the lower/upper bound of the covariates. It indicates that HBCA-N interval has the best performance among the four intervals. Meanwhile, ACNA intervals perform well too, especially when sample size gets bigger. The HAC-N intervals perform better than the Wilson score-based HWS-N intervals. When the sample size increases, the coverage probabilities of HAC-N intervals are closer to the nominal confidence level. When the errors ϵ_t's distributions are not normal, following the scaled student t-distribution $t_4/\sqrt{2}$ with variance one, from Figure 2.5, we observe that the coverage probabilities of HBCA-E and ACNA intervals for the covariate-adjusted YI are farther from the 95% nominal level at some values of covariates than the other two intervals. So we don't recommend them for the covariate-adjusted YI under the heteroscedastic regression models when errors are not normally distributed. Comparing HAC-E and HWS-E intervals, HAC-E intervals have better performance and are much more stable than HWS-E intervals in terms of coverage probability. Under the same model setting (model 2), we also examine the performance of HBCA-E and ACNA intervals for the optimal cut-off point at given $Z = z$ when the errors are not normally distributed. From Figure 2.6, we can see that although the coverage probabilities of ACNA interval are slightly higher than the 95% nominal level

at the majority values of covariates, it's still much more stable than HBCA-E. Above all, we recommend the HAC-E interval for the covariate-adjusted YI and the ACNA interval for the optimal cut-off point under the heteroscedastic regression models for the test results in practice.

2.5 Real Data Analysis

For illustrating the application of the proposed methods, we consider a dataset from the Pima Indians Diabetes Study (Smith et al., 1988). In the dataset, there are 268 cases and 500 controls. The dataset includes nine variables: the number of times pregnant (V_1), the plasma glucose concentration in an oral glucose tolerance test (OGTT) (V_2), the diastolic blood pressure (mm Hg) (V_3), the triceps skin fold thickness (mm) (V_4), 2-Hour serum insulin (mu U/ml) (V_5), body mass index (weight in kg/(height in m)^2) (V_6), diabetes pedigree function (V_7), age (years) (V_8), disease status (0 or 1) (V_9). Five observations having OGTT value 0 were deleted in the data analysis. The OGTT is a standard diagnostic test for diabetes. We want to know how accurate the OGTT is in detecting diabetes.

Based on the Pearson chi-square test for normality with p-value being 0.001 and 0.023 for cases and controls respectively, we conclude that the OGTT results from the case and the control groups are not normally distributed. The empirical estimate for the YI of the OGTT without covariate adjustment is $\widehat{J}_E = 0.446$. It indicates that the diagnostic accuracy of the OGTT is mediocre in detecting diabetes.

Smith and Thompson (1996) considered the age as a potential covariate that could influence the outcomes of the OGTT. It is interesting to know the effect of age on estimating the YI. The scatter plots of the OGTT results vs. age among case and control groups (see Figure 2.7) show that the linear regression models cannot be directly applied to model the OGTT results here. However, the heteroscedastic regression models (2.14) and (2.15) could be used to model this dataset. Figure 2.8 presents the local linear regression estimates for the mean and variance functions for both cases and controls.

Here, we use the OGTT results to obtain three estimates $\widehat{J}_N(z)$, $\widehat{J}_E(z)$ and $\widehat{J}_{AC}(z)$ with

the 95% level pointwise HBCA-N and HAC-E bands for the covariate-adjusted YI, and two estimates $\widehat{c}_{oN}(z)$ and $\widehat{c}_{oE}(z)$ with the 95% level pointwise HBCA-N and ACNA bands for the covariate-adjusted optimal cut-off point when the age z is between 21 and 66. From Figure 2.9, we observe that the diagnostic accuracy of the OGTT for younger individuals (age $<$ 30 years) is higher than that for individuals aged from 30 years to 35 years. There is a small spike which shows a slightly increasing accuracy for 38 years to 40 years old individuals, and then the accuracy decreases slowly to about 50 years. When testing individuals are getting older (age $>$ 50 years), the accuracy of OGTT increases, and the confidence bands become wider as age increases. This probably is due to the sparseness of observations when age is larger than 50. Based on our simulation studies, we would recommend the nonparametric estimate $\widehat{J}_E(z)$ and the HAC-E band for the covariate-adjusted YI to this dataset because they are more flexible and robust than the $\widehat{J}_N(z)$ estimate and the HBCA-N band which need the normal error assumption. Meanwhile, we compute the estimates $\widehat{c}_{oN}(z)$ and $\widehat{c}_{oE}(z)$ with the 95% level pointwise HBCA-N and ACNA bands for the optimal cut-off point when the age z is between 21 and 66. From Figure 2.10, we would recommend the nonparametric estimate $\widehat{c}_{oE}(z)$ and the ACNA band for the optimal cut-off point to this dataset because they are more flexible and robust than the estimate $\widehat{c}_{oN}(z)$ and the HBCA-N band which need the normal error assumption.

2.6 Discussion

Covariates are important in the evaluation of the diagnostic accuracy of a biomarker/medical test. Ignoring the covariates' effects may lead to biased estimation of the diagnostic accuracy and even wrong conclusions. Pepe (2003) gave an introduction to why and how to adjust for covariates in ROC analysis. Pardo-Fernandez et al. (2013) gave an excellent review on ROC curve analysis in the presence of covariates. One important approach to incorporate covariates to the ROC analysis is through regression models. In a parametric framework, Faraggi (2003) used simple linear regression models for the conditional means with normal errors, in both non-diseased and diseased populations, and provided a simple

method for inferences on covariate-adjusted ROC curve. It is well known that the normal distribution assumption for test results plays an important role in parametric ROC curve analysis, and the GPQ-based methods can provide 'exact' interval estimation for the Youden index under normal models for test results (see Li et al., 2008, Tian, 2011). In this paper, we have proposed the GPQ-based interval for the covariate-adjusted Youden index under linear regression models with the normal error distribution. Our simulation results have shown that the GPQ-based intervals outperform the bootstrap-based BCa, BTI and BTII intervals under the same parametric linear models setting, particularly for small to moderate sized samples which are more applicable and practical in second or third phase medical diagnostic trial studies. As indicated in the Introduction, the existing method (Faraggi, 2003) for the covariate-adjusted YI is limited to the simple linear regression models and the normality assumption for the error distributions. We generalize the approach of Faraggi (2003) and propose HWS, HAC and HBCA intervals for the covariate-adjusted Youden index under the heterosedastic regression models with/without the normality assumption for the error distributions. Our simulation results have shown that the HAC-E interval outperforms other intervals for the covariate-adjusted Youden index in most cases considered here. Furthermore, ACNA interval for the covariate-adjusted optimal cut-off is much more stable than the HBCA intervals under heteroscedastic regression models. Therefore, we recommend the use of the HAC-E interval for the covariate-adjusted YI and the ACNA interval for the covariate-adjusted optimal cut-off under the heteroscedastic regression models in practice.

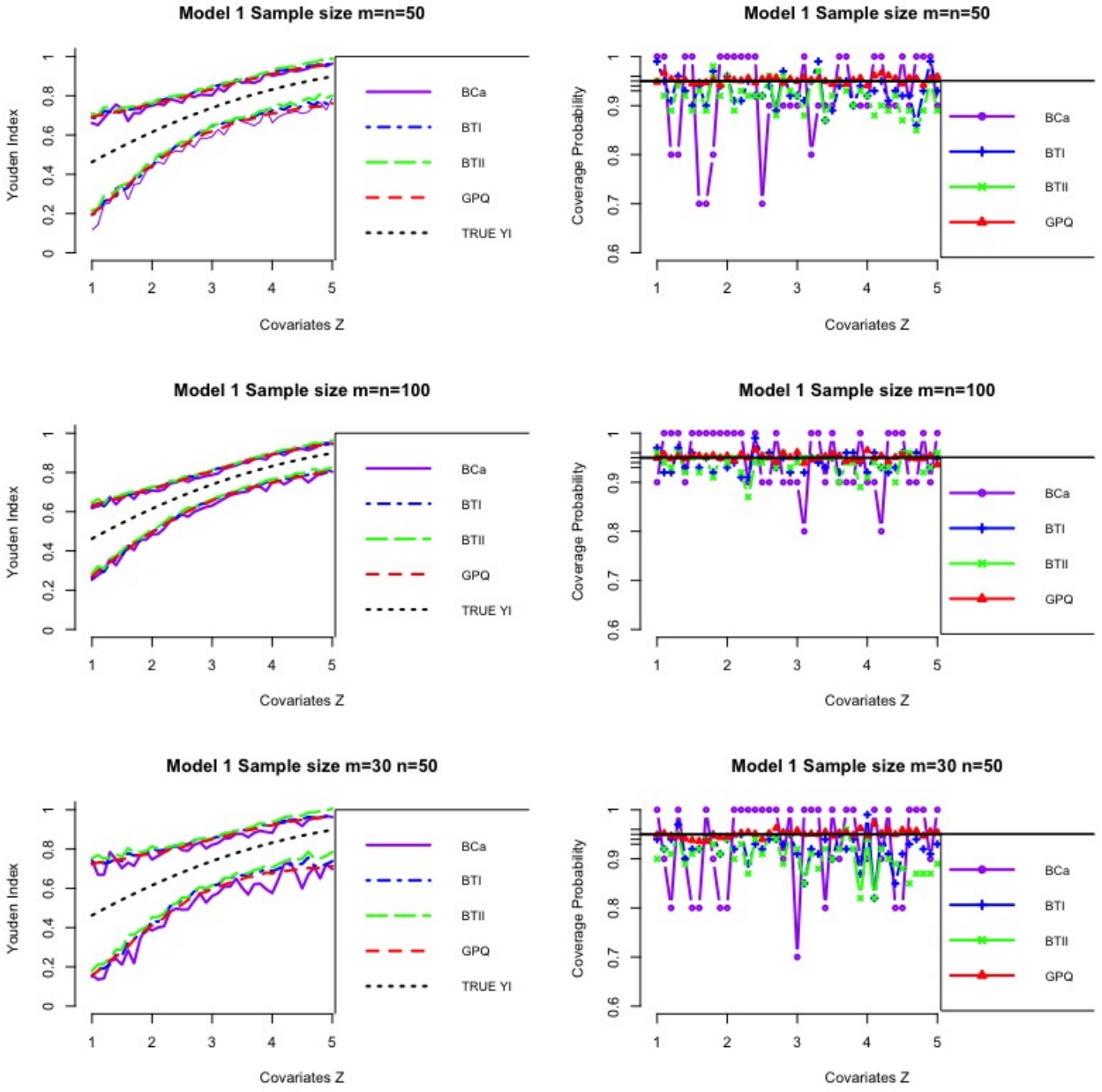

Figure (2.1) The 95% level BCa, GPQ, BTI and BTII confidence intervals for the covariate-adjusted YI at given $Z = z$ under model 1. Left panel: the 95% pointwise confidence bands for $J(z)$, the dotted curve lying in the middle is the true value of the covariate-adjusted YI. Right panel: the coverage probabilities of the BCa, GPQ, BTI and BTII intervals, solid line is the benchmark as the 95% nominal level.

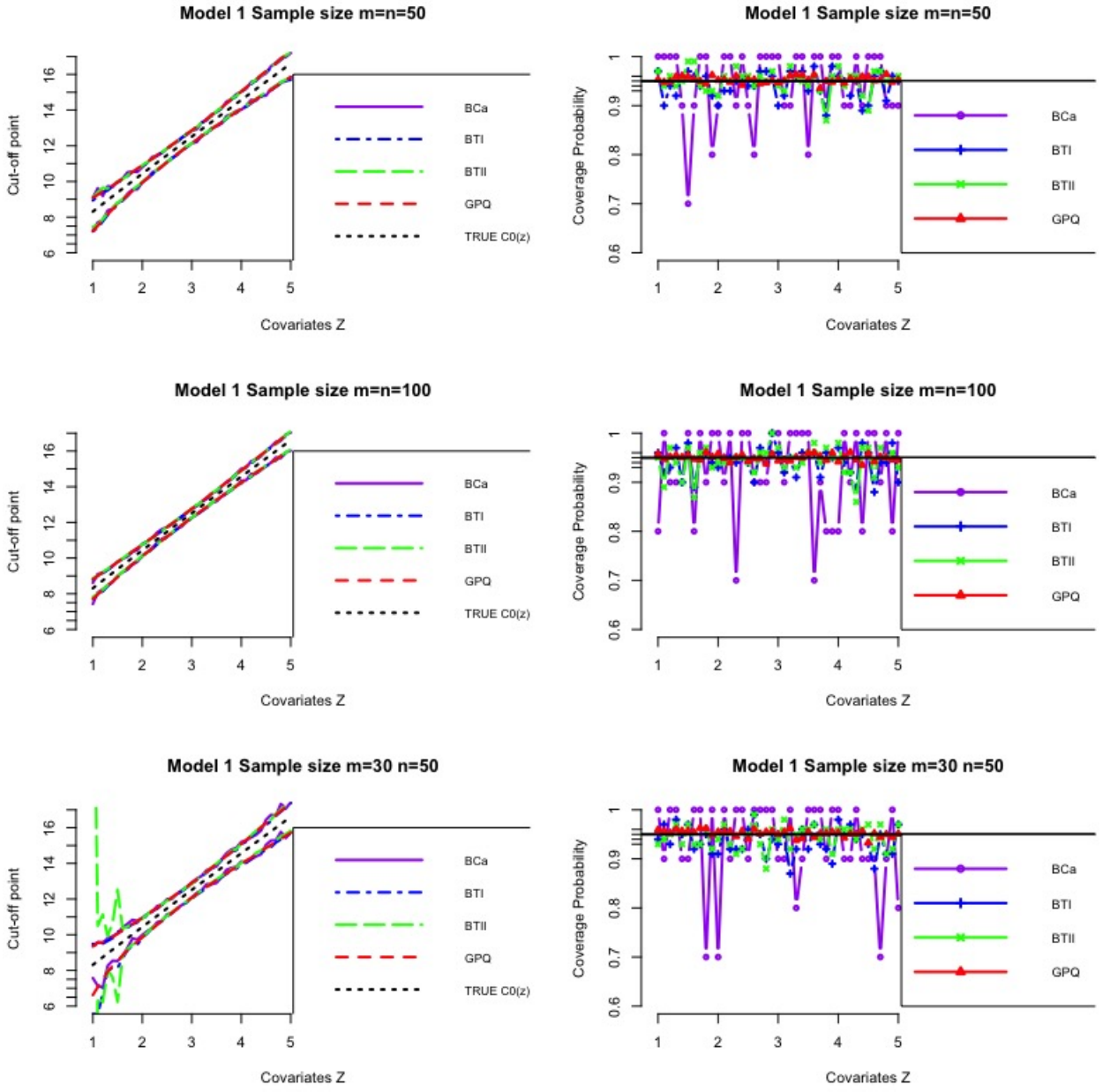

Figure (2.2) The 95% level BCa, GPQ, BTI and BTII confidence intervals for the optimal cut-off point at given $Z = z$ under model 1. Left panel: the 95% pointwise confidence bands for $c_0(z)$, the dotted curve lying in the middle is the true value of the optimal cut-off point. Right panel: the coverage probabilities of the BCa, GPQ, BTI and BTII intervals for $c_0(z)$, solid line is the benchmark as the 95% nominal level.

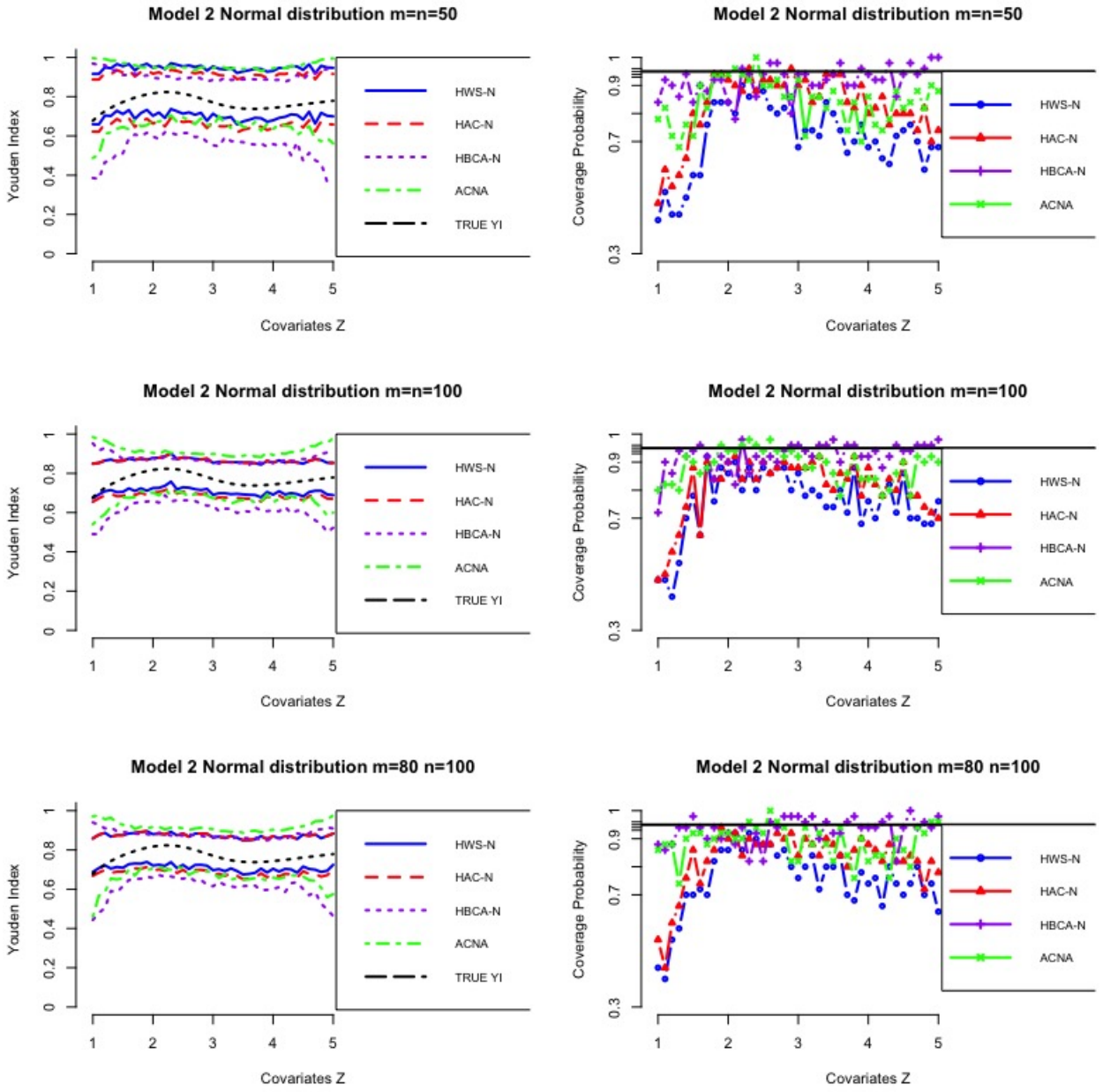

Figure (2.3) The 95% level HWS-N, HAC-N, HBCA-N and ACNA intervals for the covariate-adjusted YI at given $Z = z$ under model 2 with the normal error assumption. Left panel: The 95% pointwise confidence bands for $J_N(z)$. Right panel: The coverage probabilities of the HWS-N, HAC-N, HBCA-N and ACNA intervals for $J_N(z)$.

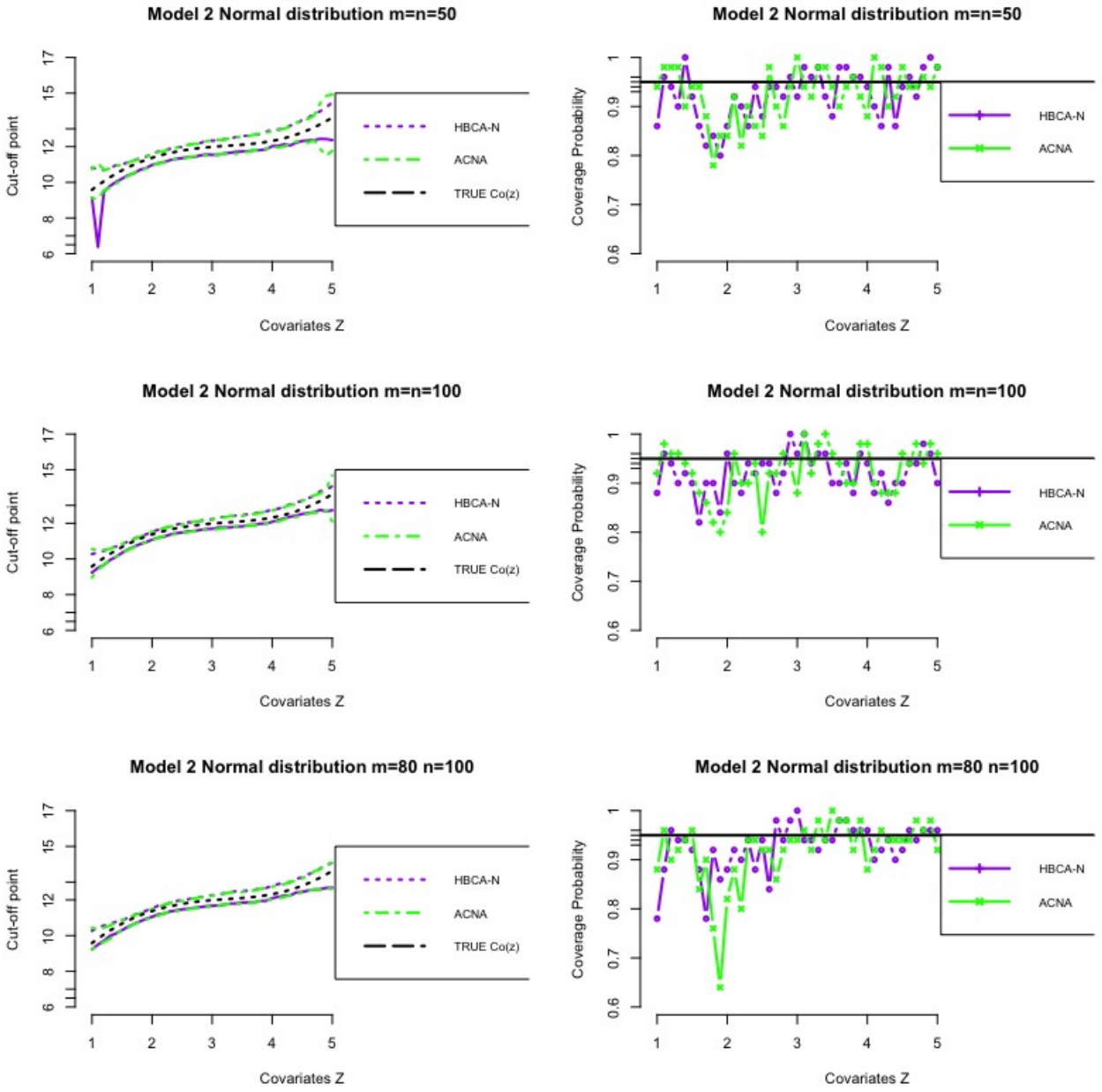

Figure (2.4) The 95% level HBCA-N and ACNA intervals for the optimal cut-off point at given $Z = z$ under model 2 with the normal error assumption. Left panel: The 95% pointwise confidence bands for $c_o(z)$. Right panel: The coverage probabilities of the HBCA-N and ACNA intervals for $c_o(z)$.

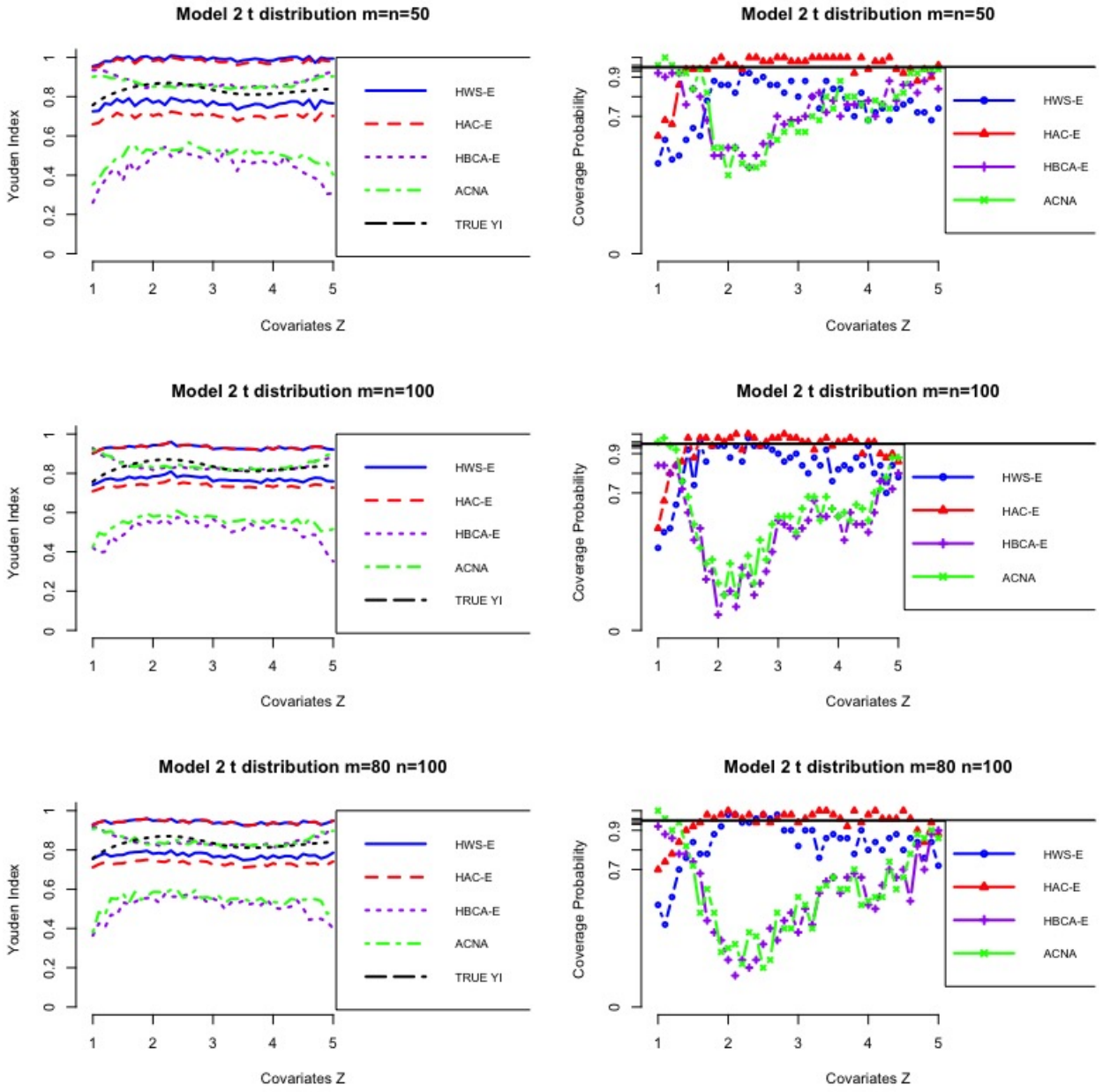

Figure (2.5) The 95% level HWS-E, HAC-E, HBCA-E and ACNA intervals for the covariate-adjusted YI at given $Z = z$ under model 2 without the normal error assumption. Left panel: The 95% pointwise confidence bands for $J(z)$. Right panel: The coverage probabilities of the HWS-E, HAC-E, HBCA-E and ACNA intervals for $J(z)$.

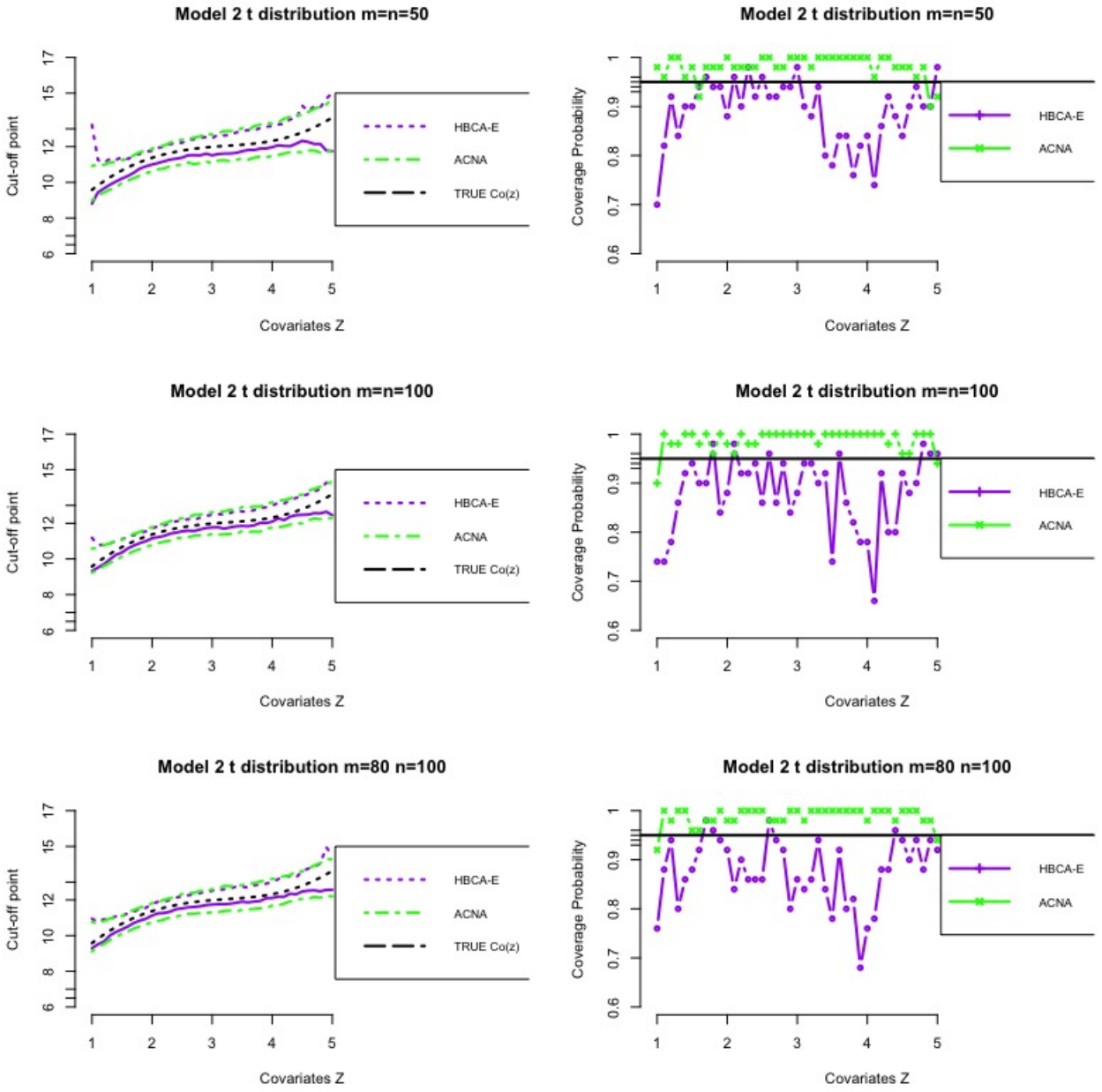

Figure (2.6) The 95% level HBCA-E and ACNA intervals for the optimal cut-off point at given $Z = z$ under model 2 without the normal error assumption. Left panel: The 95% pointwise confidence bands for $c_o(z)$. Right panel: The coverage probabilities of the HBCA-E and ACNA intervals for $c_o(z)$.

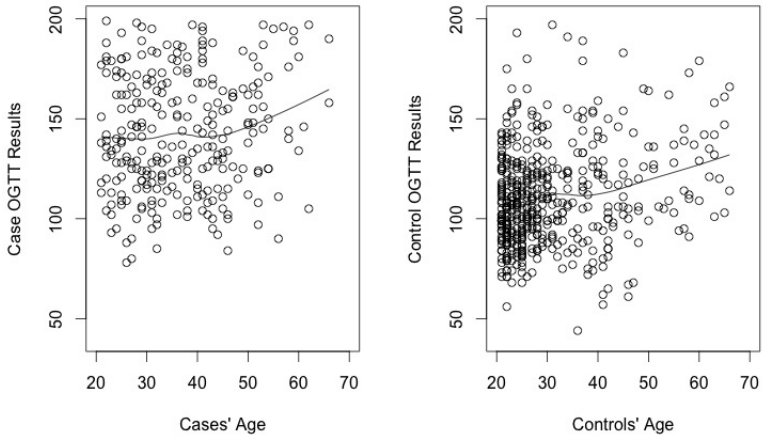

Figure (2.7) The scatter plot of OGTT test vs. Age, left for cases, right for controls. Solid lines are local linear estimates for the mean functions.

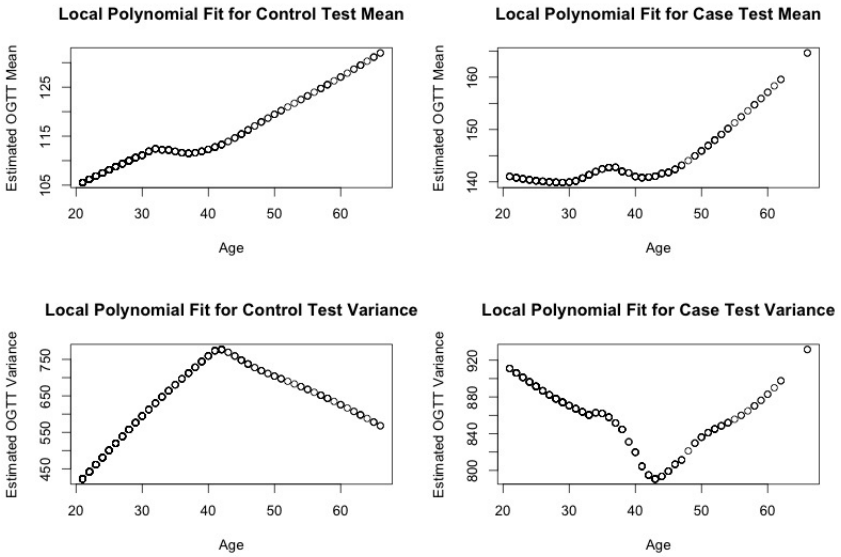

Figure (2.8) Local linear estimates for the mean and variance functions of the OGTT results from Case and Control groups.

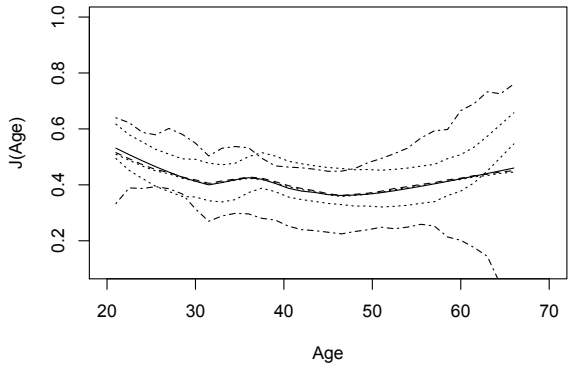

Figure (2.9) Estimates for $J(Age)$: $\widehat{J}_N$ (solid), $\widehat{J}_E$ (dashed), and $\widehat{J}_{AC}$ (dot). Point-wise confidence bands for $J(Age)$: HBCA-N (dotdash) band and HAC-E band (dot).

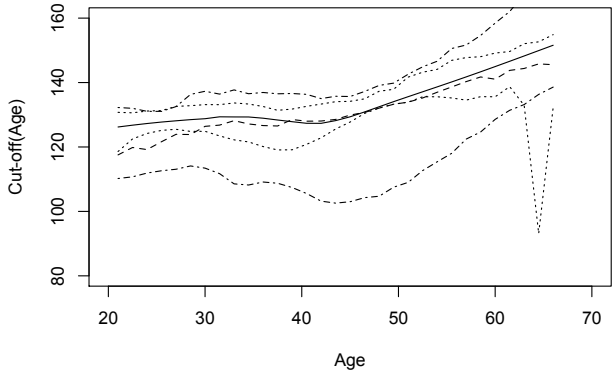

Figure (2.10) Estimates for the optimal cut-off(Age): $\widehat{c}_{oN}(z)$ (solid), $\widehat{c}_{oE}(z)$(dashed). Point-wise confidence bands for the optimal cut-off(Age): HBCA-N (dotdash) band, and ACNA band (dot).

PART 3

INFLUENCE FUNCTION-BASED EMPIRICAL LIKELIHOOD INFERENCES FOR AUC IN THE PRESENCE OF COVARIATES

3.1 Introduction of ROC Curve and AUC

In medical diagnostic studies, ROC curve stands for *Receiver Operating Characteristic* curve, which is the plot of sensitivity versus one minus specificity for all possible threshold values (See Pepe [33], Zhou *et al.* [34]). The area under the ROC curve (AUC) is a popular summary measurement of the diagnostic accuracy of a continuous scale test. A subject is classified as diseased (positive) if the subject's test value is greater than a chosen threshold value, and otherwise classified as non-diseased (negative). Let Y^D be the test result of a diseased subject and $Y^{\bar{D}}$ be the test result of a non-diseased subject. Then the AUC can be expressed as $AUC = P(Y^D > Y^{\bar{D}})$ (Bamber, 1975). Obviously the value of the AUC is between 0 and 1. The closer the ROC curve follows the left-hand border and then the top border of the ROC space, the closer to one the AUC value, and the more accurate the diagnostic test.

3.2 Motivation

Covariate-adjustments for summary measures of the ROC curve have become crucial in many diagnostic applications. Since operating conditions for the test or characteristics of patients like gender, age, race, physical conditions and so on, may affect the test results by influencing the distributions of test measurements for "diseased" and/or "non-diseased" subjects. Many researches consider incorporating covariates information when conducting the regression analysis. Thompson and Zucchini [35] and Obuchowski [36] proposed AUC regression methods based on the derived variable. Dorfman, Berbaum, and Metz [37] developed a method by computing jackknife AUC values for each subject. However, these

methods can only be applied for discrete covariates (Dodd and Pepe[38]).

Normal approximation is a commonly used method for constructing confidence intervals/regions for regression parameters. Dodd and Pepe [38] proposed a regression model for the AUC summary measure. However, the asymptotic variance of the estimator for the parametric vector is of a very complicated form and the explicit estimate for the asymptotic variance of the parameter vector was not well developed in the literature. Instead, Dodd and Pepe suggested using bootstrap method to estimate it. Our simulation studies indicated that the normal approximation-based confidence regions for β have poor coverage accuracy by using the bootstrap variance estimate. To overcome these problems, in this part, we will apply two empirical likelihood based methods to construct confidence regions for β: one is an empirical likelihood confidence region based on influence function, the other is a Jackknife empirical likelihood-based confidence region. The proposed methods allow for confidence region construction without a variance estimator. Our simulation study shows that the proposed methods have better small sample performances than the existing normal approximation-based method in terms of coverage probability.

3.3 Normal Approximation-based Method

Let $\theta_{ij} = P(Y_j^D > Y_i^{\bar{D}} | \mathbf{Z_i^{\bar{D}}}, \mathbf{Z_j^D})$ be the covariate-specific AUC parameter. Dodd and Pepe [38] defined the AUC regression model as

$$\theta_{ij} = g(\beta^T \mathbf{Z_{ij}}),$$

where $\mathbf{Z_{ij}}$ denote the observable covariates $(\mathbf{Z_i^{\bar{D}}}, \mathbf{Z_j^D})$, g is a specified function. Denote $I_{ij} = I(Y_j^D > Y_i^{\bar{D}})$. In order to obtain an estimator for β, Dodd and Pepe [38] proposed the following generalized estimating equation:

$$\sum_{j=1}^{n_D} \sum_{i=1}^{n_{\bar{D}}} \frac{\partial \theta_{ij}}{\partial \beta} \omega(\mathbf{Z_{ij}}, \beta)(\mathbf{I_{ij}} - \mathbf{g}(\beta^{\mathbf{T}} \mathbf{Z_{ij}})) = \mathbf{0}, \tag{3.1}$$

where $N = n_D + n_{\bar{D}}$ and $\omega(\mathbf{Z_{ij}}, \beta)$ is a known weight function. They obtained the estimator $\hat{\beta}$ for β from the estimating equation and showed that the distribution of the estimator is asymptotically normal:

$$\sqrt{\frac{n_{\bar{D}} n_D}{N}}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} N(0, \Xi), \quad \text{as} \quad N \to \infty, \tag{3.2}$$

where Ξ is the asymptotic variance of $\widehat{\beta}$.

We can observe that if a good estimate for Ξ can be obtained, this asymptotic normal distribution could be used to construct a confidence region for β. However, the asymptotic variance Ξ in (3.2) is of a very complicated form. Dodd and Pepe suggested using bootstrap method to estimate Ξ, and a normal approximation-based confidence region (NA) for β can be constructed as follows:

$$CR_{NA}(\beta) = \left\{ \beta : \frac{N}{n_{\bar{D}} n_D} \cdot (\widehat{\beta} - \beta_0)^T \Xi^{*-1} (\widehat{\beta} - \beta_0) \leq \chi^2_{p,1-\alpha} \right\} \tag{3.3}$$

where Ξ^* is the bootstrap estimate for the asymptotic variance of $\widehat{\beta}$.

3.4 Influence Function-based Empirical Likelihood for the AUC Regression

Empirical Likelihood (EL) is a non-parametric method introduced by Owen [12][13]. EL-based methods have been successfully applied to ROC analysis (Claeskens *et al.*[16], Qin and Zhou [17]).

Let $\{Y_i^{\bar{D}},\ i = 1 \ldots m\}$ denote a sample of test results from non-diseased subjects following the distribution function $F^{\bar{D}}_{\mathbf{Z^{\bar{D}}}}$ with covariate vectors $\mathbf{Z}_i^{\bar{D}} = (Z_{i1}^{\bar{D}}, \cdots, Z_{iq}^{\bar{D}})^T$, and let $\{Y_j^D,\ j = 1 \ldots n\}$ denote a sample of test results from diseased subjects following the distribution function $F^D_{\mathbf{Z^D}}$ with covariate vectors $\mathbf{Z}_j^D = (Z_{j1}^D, \cdots, Z_{jp}^D)^T$. Let Y be the test result of the continuous-scale diagnosis. Then, we can define the placement value of Y as $U = 1 - F^{\bar{D}}_{\mathbf{Z^{\bar{D}}}}(Y)$. Particularly, $U^{\bar{D}} = 1 - F^{\bar{D}}_{\mathbf{Z^{\bar{D}}}}(Y^{\bar{D}})$ and $U^D = 1 - F^{\bar{D}}_{\bar{D}}(Y^D)$, where $U^{\bar{D}}$ follows uniform distribution in $(0, 1)$ and U^D measures the discrimination be-

tween the diseased and non-diseased subjects. The conditional ROC curve is represented by $ROC_{\mathbf{Z}}(u) = P(U^D < u|\mathbf{Z}^D)$, where $u = 1 - F^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(y)$ is a false positive rate. The covariate-specific AUC can be expressed as a conditional expectation function of $1 - U^D$ given $\mathbf{Z}^D$, which is $AUC_{\mathbf{Z}} = E(1 - U^D|\mathbf{Z}^D)$. The covariate effects on discrimination can be evaluated by using the AUC regression models (Pepe and Cai, [39]; Dodd and Pepe, [33]). Here, we consider the following AUC regression model:

$$AUC_{\mathbf{Z}} = E(1 - U^D|\mathbf{Z^D}) = g(\beta^T\mathbf{Z}^D), \tag{3.4}$$

where $\mathbf{Z}^D = (Z_1^D, \cdots, Z_p^D)^T$ is the $p \times 1$ vector of covariates, g is a specified link function.

To estimate β, we use the GLM re-weighted least squares fitting-based estimation equation:

$$\sum_{j=1}^{n} \omega(\beta^T\mathbf{Z}_j^D)[1 - U_j^D - g(\beta^T\mathbf{Z}_j^D)]\mathbf{Z}_j^D \equiv \sum_{j=1}^{n} \mathbf{H}_j = \mathbf{0}, \tag{3.5}$$

where $\mathbf{H}_j = (1-U_j^D-g(\beta^T\mathbf{Z}_j^D))\omega(\beta^T\mathbf{Z_j^D})\mathbf{Z_j^D}$, and $\omega(\beta^T\mathbf{Z}_j^D)$ is a given scalar weight function, $\mathbf{0} = (0, \cdots, 0)^T$ is a $p\times 1$ vector and $\mathbf{Z}_j^D$ is the $p\times 1$ vector of covariates for the $j-$th diseased subject.

Since the distribution for the test result in the non-diseased population is unknown, the placement value $U^D = 1 - F^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(Y^D)$ is unobservable. However, using the non-diseased sample $\{(Y_i^{\bar{D}}, \mathbf{Z}_i^{\bar{D}}), i = 1, 2, \cdots, m\}$, we can estimate the reference distribution $F^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}$ by its empirical distribution $\widehat{F}^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}$. Hence, the placement value $U_j^D = 1-F^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(Y_j^D)$ can be estimated by $\widehat{U}_j^D = 1 - \widehat{F}^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(Y_j^D)$, and $\widehat{\beta}$, an estimator for β, can be found by solving the following estimation equation:

$$\sum_{j=1}^{n} \omega(\beta^T\mathbf{Z}_j^D)[1 - \widehat{U}_j^D - g(\beta^T\mathbf{Z}_j^D)]\mathbf{Z}_j^D \equiv \sum_{j=1}^{n} \widehat{\mathbf{H}}_j = \mathbf{0}, \tag{3.6}$$

where $\widehat{\mathbf{H}}_j = (1 - \widehat{U}_j^D - g(\beta^T\mathbf{Z}_j^D))\omega(\beta^T\mathbf{Z_j^D})\mathbf{Z_j^D}$.

Let $\{(Y_k, \mathbf{Z_k}), k = 1, \cdots, n+m\} = \{(Y_1^D, \mathbf{Z_1^D}), \cdots, (Y_n^D, \mathbf{Z_n^D}), (Y_1^{\bar{D}}, \mathbf{Z_1^{\bar{D}}}), \cdots, (\mathbf{Y_m^{\bar{D}}}, \mathbf{Z_m^{\bar{D}}})\}$.

From the proof of Lemma 3.1 showed in Appendix A, we obtain that

$$\begin{aligned}
\frac{1}{\sqrt{n+m}}\sum_{j=1}^{n}\widehat{\mathbf{H}}_j &= \frac{1}{\sqrt{n+m}}\sum_{j=1}^{n}\mathbf{H_j}+\frac{n}{m}[\lim_n\frac{1}{n}\sum_{j=1}^{n}\omega(\beta^T\mathbf{Z}_j^D)\mathbf{Z}_j^D] \\
&\quad\times\frac{1}{\sqrt{n+m}}\sum_{i=1}^{m}\left(\int I(Y_i^{\bar{D}}\le t)dF_{\mathbf{Z^{\bar{D}}}}^D(t)-\int F_{\mathbf{Z^{\bar{D}}}}^{\bar{D}}(t)dF_{\mathbf{Z^D}}^D(t)\right)+o_p(1) \\
&= \frac{1}{\sqrt{n+m}}\sum_{j=1}^{n}\mathbf{H_j}+\frac{\rho A(\beta)}{\sqrt{n+m}}\sum_{i=1}^{m}\int\left(I(Y_i^{\bar{D}}\le t)-F_{\mathbf{Z^{\bar{D}}}}^{\bar{D}}(t)\right)dF_{\mathbf{Z^D}}^D(t)+o_p(1) \\
&\equiv \frac{1}{\sqrt{n+m}}\sum_{k=1}^{n+m}W_k(\beta)+o_p(1),
\end{aligned}$$

where $\rho=\lim_n\frac{n}{m}$, $A(\beta)=\lim_n\frac{1}{n}\sum_{j=1}^{n}\omega(\beta^T\mathbf{Z}_j^D)\mathbf{Z}_j^D$, and

$$W_k(\beta)=\begin{cases}\mathbf{H_k}, & \text{if } k=1,\cdots,n,\\ \rho A(\beta)\int(I(Y_k\le t)-F_{\mathbf{Z^{\bar{D}}}}^{\bar{D}}(t))dF_{\mathbf{Z^D}}^D(t), & \text{if } k=n+1,\cdots,n+m.\end{cases}$$

is the k-th influence function of β.

Using the side information provided by the influence function, the empirical likelihood for β can be defined as follows

$$L(\beta)=\sup\left\{\prod_{k=1}^{n+m}p_k : p_1\ge 0,\cdots,p_{n+m}\ge 0,\sum_{k=1}^{n+m}p_k=1,\sum_{k=1}^{n+m}p_k\widehat{W}_k(\beta)=\mathbf{0}\right\}, \tag{3.7}$$

where

$$\widehat{W}_k(\beta)=\begin{cases}\widehat{\mathbf{H}}_\mathbf{k}, & \text{if } k=1,\cdots,n,\\ \frac{1}{m}\sum_{j=1}^{n}\omega(\beta^T\mathbf{Z}_j^D)\mathbf{Z}_j^D\int(I(Y_k\le t)-\widehat{F}_{\mathbf{Z^{\bar{D}}}}^{\bar{D}}(t))d\widehat{F}_{\mathbf{Z^D}}^D(t), & \text{if } k=n+1,\cdots,n+m.\end{cases}$$

is the k-th estimated influence function, $\widehat{F}_{\mathbf{Z^{\bar{D}}}}^{\bar{D}}(t)=\frac{1}{m}\sum_{i=1}^{m}I(Y_i^{\bar{D}}\le t)$, and $\widehat{F}_{\mathbf{Z^D}}^D(t)=\frac{1}{n}\sum_{j=1}^{n}I(Y_j^D\le t)$.

Using the Lagrange multiplier method, we get that

$$p_k=\frac{1}{n+m}\frac{1}{1+\nu^T\widehat{W}_k(\beta)},$$

where $\nu^T = (\nu_1, \nu_2, ..., \nu_p)$ is the solution to

$$\frac{1}{n+m}\sum_{k=1}^{n+m}\frac{\widehat{W}_k(\beta)}{1+\nu^T\widehat{W}_k(\beta)} = \mathbf{0}. \tag{3.8}$$

Note that $\prod_{k=1}^{n+m} p_k$, subject to $\sum_{k=1}^{n+m} p_k = 1, p_k \geq 0, k = 1, 2, ..., n+m$, attains its maximum $(n+m)^{-(n+m)}$ at $p_k = (n+m)^{-1}$. So, the influence function-based empirical likelihood ratio for β is

$$R(\beta) = \prod_{k=1}^{n+m}(n+m)p_k = \prod_{k=1}^{n+m}\{1+\nu^T\widehat{W}_k(\beta)\}^{-1}.$$

The corresponding influence function-based empirical log-likelihood ratio for β is

$$l(\beta) = -2\log R(\beta) = 2\sum_{k=1}^{n+m}\log(1+\nu^T\widehat{W}_k(\beta)). \tag{3.9}$$

Theorem 3.1. *If* $max_j\|\mathbf{Z_j^D}\| = o_p(n^{1/2})$, $\lim_{n,m\to\infty}\frac{n}{m} = \rho > 0$, *g and* ω *are bounded functions, and* β_0 *is the true parameter vector in the AUC regression model, then the asymptotic distribution of the influence function-based empirical log-likelihood ratio statistic* $l(\beta_0)$ *is a chi-square distribution with p degree of freedom. That is,*

$$l(\beta_0) \xrightarrow{\mathcal{L}} \chi_p^2. \tag{3.10}$$

A $(1-\alpha)$ level influence function-based confidence region (IFEL region) for β is

$$CR(\beta) = \{\beta : l(\beta) \leq \chi_{p,\alpha}^2\},$$

where $\chi_{p,\alpha}^2$ is the $(1-\alpha)$-th quantile of χ_p^2. By Theorem 3.1, we have that

$$P\{\beta_0 \in CR(\beta)\} = P\{l(\beta) \leq \chi_{p,\alpha}^2\} = 1-\alpha+o(1).$$

3.5 Jackknife Empirical Likelihood Method for the AUC Regression

Jackknife empirical likelihood (JEL), proposed by Jing, Yuan, and Zhou [40], is a powerful non-parametric method to overcome the computational difficulties dealing with nonlinear functionals, with the particular application to U-statistics. The JEL combines jackknife and empirical likelihood methods. Using the techniques in Jing, Yuan and Zhou [40], we can define the JEL for β. Let

$$T_{n+m}(\beta) = \frac{1}{n+m}\sum_{k=1}^{n+m}\widehat{W}_k(\beta),$$

$$T_{n+m,-i}(\beta) = \frac{1}{n+m-1}\sum_{k=1,k\neq i}^{n+m}\widehat{W}_{k,-i}(\beta),\ i=1,2,\cdots,n+m,$$

where $\widehat{W}_{k,-i}(\beta)$ is the $\widehat{W}_k(\beta)$ based on the $n+m-1$ observations from $\{(Y_k,\mathbf{Z_k}), k = 1,\cdots,n+m\}$ by deleting the i-th observation $(Y_i,\mathbf{Z_i})$. Then the jackknife pseudo sample can be written as:

$$W_i(\beta) = (n+m)T_{n+m}(\beta) - (n+m-1)T_{n+m,-i}(\beta),\ i=1,2,\cdots,n+m.$$

Applying Owen's EL to this jackknife pseudo sample, we get the following JEL for β:

$$L_J(\beta) = \sup\left\{\prod_{k=1}^{n+m}p_k : p_1\geq 0,\cdots,p_{n+m}\geq 0, \sum_{k=1}^{n+m}p_k = 1, \sum_{k=1}^{n+m}p_kW_k(\beta) = \mathbf{0}\right\} \tag{3.11}$$

The corresponding jackknife empirical log-likelihood ratio for β is

$$l_J(\beta) = 2\sum_{k=1}^{n+m}\log(1+\nu_J^TW_k(\beta)).$$

where $\nu_J^T = (\nu_{J1},\nu_{J2},...,\nu_{Jp})$ is the solution to

$$\frac{1}{n+m}\sum_{k=1}^{n+m}\frac{W_k(\beta)}{1+\nu_J^TW_k(\beta)} = \mathbf{0}. \tag{3.12}$$

Using the method similar to that in Jing, Yuan and Zhou [40], it can be proved that the asymptotic distribution of the empirical log-likelihood ratio statistic $l_J(\beta_0)$ is a chi-square distributions with p degree of freedom. That is,

$$l_J(\beta_0) \xrightarrow{\mathcal{L}} \chi^2_p. \tag{3.13}$$

Hence, a $(1-\alpha)$ level JEL-based confidence region (JEL region) for β can be constructed as follows

$$CR_J(\beta) = \{\beta : l_J(\beta) \leq \chi^2_{p,\alpha}\},$$

where $\chi^2_{p,\alpha}$ is the $(1-\alpha)$-th quantile of χ^2_p.

3.6 Empirical Likelihood-based Confidence Interval for AUC_Z

Pepe and Cai [39] proposed a pseudo likelihood-based inference in ROC regression model. In this section, we propose an EL-based confidence interval for covariate-specific AUC in the AUC regression model.

Let $CR(\beta)$ be a $(1-\alpha)$-th confidence interval for β. Then we can construct a $(1-\alpha)$-th confidence interval for the covariate-specific AUC given a specified covariate $\mathbf{Z^D}$ as follows:

$$\{AUC_Z(\beta) = g(\beta^T \mathbf{Z^D}) : \beta \in \mathbf{CR}(\beta)\}, \tag{3.14}$$

where $g(\cdot)$ is a one-to-one function.

Let (q_0, q_1) denote the confidence interval for the covariate adjusted AUC_Z. In order to

compute the confidence interval (see also Zhou *et al.*, [41]), we apply the following equations:

$$\begin{aligned} q_0 &= min\{AUC_Z(\beta) : \beta \in CR(\beta)\} = min\{AUC_Z(\beta) : l_1(\beta) = c, 0 \le c \le c_\alpha\} \\ &\approx min\{\bigcup_{i=1}^{N}(AUC_Z(\beta) : l_1(\beta) = c_i)\}, \\ q_1 &= max\{AUC_Z(\beta) : \beta \in CR(\beta)\} = max\{AUC_Z(\beta) : l_1(\beta) = c, 0 \le c \le c_\alpha\} \\ &\approx max\{\bigcup_{i=1}^{N}(AUC_Z(\beta) : l_1(\beta) = c_i)\}, \end{aligned}$$

where N is a large integer number, $\{c_1, c_2, ..., c_N\}$ is a random sample of size N generated from the uniform distribution on $[0, c_\alpha]$.

In order to estimate q_0, q_1, we use the following approximation procedure:

1. For $b = 1, 2, ..., B$, where B is a chosen integer depending on the number of regression parameters, we generate B vectors for $\{\beta^{(b)}\}$ uniformly over CR_0 satisfying $l_1(\beta^{(b)}) \le c_\alpha$ for $b = 1, 2, \cdots, B$, by smoothing technique (e.g., the local linear method).

2. We approximate $CR(\beta)$ by $CR_0 = \{\beta : \widehat{\beta}_k - z_{1-\alpha/2}\widehat{\sigma}_k \le \beta_k \le \widehat{\beta}_k + z_{1-\alpha/2}\widehat{\sigma}_k,\ k = 1, \cdots, p\}$, where $\widehat{\sigma}_k$ is the standard error of $\widehat{\beta}_k$ and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th quantile of the standard normal distribution.

3.7 Simulation Study

In this section, we conduct two simulation studies to evaluate finite sample performances of the proposed EL-based methods in the AUC regression. Particularly, we compare the influence function-based IFEL confidence region, the Jackknife-based JEL confidence region with the existing normal approximation-based NA confidence region for β in the AUC regression model in terms of coverage probability. We use the similar AUC regression simulation setting illustrated in Dodd and Pepe [38].

Let $Y^D|\mathbf{Z^D} \sim \mathbf{N}(\mu_{\mathbf{D},\mathbf{Z^D}}, \sigma^{\mathbf{2}}_{\mathbf{D}})$ and $Y^{\bar{D}}|\mathbf{Z^{\bar{D}}} \sim \mathbf{N}(\mu_{\mathbf{\bar{D}},\mathbf{Z^{\bar{D}}}}, \sigma^{\mathbf{2}}_{\mathbf{\bar{D}}})$. Then, we have

$$AUC_{\mathbf{Z}} = \Phi\left(\frac{\mu_{D,\mathbf{Z^D}} - \mu_{\bar{D},\mathbf{Z^{\bar{D}}}}}{\sqrt{\sigma^2_D + \sigma^2_{\bar{D}}}}\right).$$

We carry out simulation studies with sample size $m, n = 30$, 50, 100 for both diseased group and non-diseased group, respectively. By generating 1000 random samples based on the above simulation setting and applying the usual generalized linear regression method (GLM), we can obtain estimate $\widehat{\beta}$ for β and calculate the coverage probabilities of the NA-based confidence region, JEL confidence region and the proposed IFEL confidence region for β.

In the study, we choose dimension $p = 2, 3$ for first and second model, respectively. The covariates $\mathbf{Z^D}$ and $\mathbf{Z^{\bar{D}}}$ are assumed to have common components in both models.

Model 1 ($p = 2$):

$$\mu_{\bar{D},Z} = \gamma_0 + \gamma_1 Z,$$

$$\mu_{D,Z} = k_0 + k_1 Z.$$

The first model is based on the single covariate Z, where the common covariate Z follows $U(0, 10)$. We choose $\gamma_0 = 0$, $\gamma_1 = 0$, $k_0 = 0$, $k_1 = 0.5$. Then the AUC regression model is

$$AUC_Z = \Phi\left(\frac{(k_0 - \gamma_0) + (k_1 - \gamma_1)Z}{\sqrt{\sigma^2_D + \sigma^2_{\bar{D}}}}\right) = \Phi(\beta_0 + \beta_1 Z), \tag{3.15}$$

where we choose $\sigma_D = 1.2$, $\sigma_{\bar{D}} = 1$ so that the true parameters $\beta_0 = (k_0 - \gamma_0)/\sqrt{\sigma^2_D + \sigma^2_{\bar{D}}} = 0$, $\beta_1 = (k_1 - \gamma_1)/\sqrt{\sigma^2_D + \sigma^2_{\bar{D}}} = 0.32$.

From our simulation results (not reported here), the IFEL region is too conservative (with coverage probability close to 1) with small sample sizes. In order to get an IFEL-based confidence region with better coverage accuracy, we propose the following bootstrap

procedure to construct an IFEL-based region for β .

1. Draw a bootstrap resample $\{Y_i^{*\bar{D}} : i = 1, \cdots, m\}$ from the "non-diseased" sample $\{Y_i^{\bar{D}} : i = 1, \cdots, m\}$, and a bootstrap resample $\{Y_j^{*D} : j = 1, \cdots, n\}$ from the "diseased" sample $\{Y_j^D : j = 1, \cdots, n\}$, respectively.

2. Applying GLM method to get the estimate $\widehat{\beta}$ for β and compute the bootstrap copy $\widehat{W}_k^*(\widehat{\beta})$ of $\widehat{W}_k(\widehat{\beta})$.

3. Repeat the first two steps B times to obtain B bootstrap copies $\{l^{*b}(\widehat{\beta}),\ b = 1, \cdots, B\}$ of $l(\beta)$ by (3.9).

Then $(1-\alpha)100\%$ $(0 < \alpha < 1)$ level bootstrap-based IFEL (BIFEL) region for β can be constructed as follows:

$$\{\beta : l(\beta) \in (l(\widehat{\beta})^{*([B\alpha/2])}, l(\widehat{\beta})^{*([B(1-\alpha/2)])})\},$$

where $l(\beta)^{*([B\alpha/2])}$ and $l(\beta)^{*([B(1-\alpha/2)])}$ are the $\alpha/2$-th and $(1-\alpha/2)$-th quantiles of $\{l^{*b}(\widehat{\beta}),\ b = 1, \cdots, B\}$, respectively.

The parameter estimates and the coverage probabilities for β are presented in Tables 3.1 - 3.2. From Table 3.1, we can see that the GLM method provides good estimates for β.

Table (3.1) The parameters estimates in the AUC regression model $AUC_{\mathbf{Z}} = \Phi(\beta_0 + \beta_1 Z_1)$

n	m	β_0	$\hat{\beta}_0$	bias of $\hat{\beta}_0$	sd of $\hat{\beta}_0$
30	30	0	-0.0963899	-0.0963899	0.3710366
50	50	0	-0.0340966	-0.0340966	0.2919486
100	100	0	-0.0001580	-0.0001580	0.2245143
n	m	β_1	$\hat{\beta}_1$	bias of $\hat{\beta}_1$	sd of $\hat{\beta}_1$
30	30	0.3201	0.3567006	0.0366006	0.1094567
50	50	0.3201	0.3388887	0.0187887	0.0827232
100	100	0.3201	0.3289982	0.0088982	0.0559346

From Table 3.2, we observe that the BIFEL confidence region performs the best among three confidence regions for β. Particularly, when sample size gets bigger, coverage probabilities of the BIFEL confidence regions are closer to the nominal levels. JEL performs well too. When sample size is small $(m, n = 30)$, the NA confidence regions have over-coverage problems for the true regression parameters.

Model 2 ($p = 3$):

$$\mu_{\bar{D},Z} = \gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2,$$

$$\mu_{D,Z} = k_0 + k_1 Z_1 + k_2 Z_2.$$

In the second model, covariates Z_1 and Z_2 follow $Z_1 \sim U(0, 10)$ and $Z_2 \sim N(1.2, 3)$, respectively. We choose $\gamma_0 = 0.7$, $\gamma_1 = -0.3$, $\gamma_2 = 0.3$, $k_0 = 1$, $k_1 = -0.8$, $k_2 = 1.7$. Then the AUC regression model is $AUC_{\mathbf{Z}} = \Phi(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2)$ with true parameter values $\beta_0 = 0.192$, $\beta_1 = -0.32$, and $\beta_2 = 0.896$ when $\sigma_D = 1.2$, and $\sigma_{\bar{D}} = 1$.

Based on this simulation setting, we do similar computation to obtain parametric estimates for β and coverage probabilities of confidence regions for β. The results are presented in Tables 3.3 - 3.4.

From Table 3.3, we can see that the estimates for β_0 have large biases and standard errors compared to the estimates in model 1, although the GLM method provides acceptable estimates for β_1 and β_2. It's possibly due to the true placement value $U^{\bar{D}}$ is unobservable.

Table (3.2) Coverage probabilities of 90% and 95% confidence regions for the parameters vector in the AUC regression model $AUC_{\mathbf{Z}} = \Phi(\beta_0 + \beta_1 Z_1)$

Level	n	m	BIFEL	JEL	NA
95%	30	30	0.96	0.96	0.98
	50	50	0.91	0.94	0.96
	100	100	0.96	0.92	0.94
90%	30	30	0.93	0.90	0.96
	50	50	0.85	0.84	0.93
	100	100	0.90	0.87	0.89

Table (3.3) The parameters estimates in the AUC regression model $AUC_{\mathbf{Z}} = \Phi(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2)$

n	m	β_0	$\hat{\beta}_0$	bias of $\hat{\beta}_0$	sd of $\hat{\beta}_0$
30	30	0.19205	0.8724950	0.6804450	1.0284150
50	50	0.19205	0.9224847	0.7304347	0.7748316
100	100	0.19205	0.9207765	0.7287265	0.6470202
n	m	β_1	$\hat{\beta}_1$	bias of $\hat{\beta}_1$	sd of $\hat{\beta}_1$
30	30	-0.32009	-0.5046025	-0.1845125	0.3076757
50	50	-0.32009	-0.5021865	-0.1820965	0.2288338
100	100	-0.32009	-0.5162207	-0.1961307	0.2067056
n	m	β_2	$\hat{\beta}_2$	bias of $\hat{\beta}_2$	sd of $\hat{\beta}_2$
30	30	0.89625	1.094424	0.1981746	0.4978437
50	50	0.89625	1.061623	0.165373	0.3940743
100	100	0.89625	1.104876	0.208626	0.3537151

Table (3.4) Coverage probabilities of 90% and 95% confidence regions for the parameters vector in the AUC regression model $AUC_{\mathbf{Z}} = \Phi(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2)$

Level	n	m	BIFEL	JEL	NA
95%	30	30	0.89	0.85	0.86
	50	50	0.88	0.81	0.69
	100	100	0.92	0.83	0.31
90%	30	30	0.86	0.78	0.79
	50	50	0.84	0.75	0.56
	100	100	0.88	0.74	0.21

Table 3.4 indicates that the coverage probabilities of the NA confidence regions are far below the nominal confidence levels. One possible reason for the poor performances of the NA method is that the NA-based method is sensitive to the poor estimate for the asymptotic variance of $\widehat{\beta}$. JEL regions have under-coverage problems for all the sample sizes. However, the BIFEL confidence regions have much better coverage accuracy than the NA confidence regions and JEL confidence regions, particularly for large sample sizes. BIFEL method is more robust and accurate than the NA-based method and works better than the JEL method as well. The conclusion that BIFEL confidence region outperforms the other two confidence regions is consistent with that drawn from the simulation study under model 1.

3.8 A Real Example

In this section, we apply a study of the distortion product otoacoustic emissions (DPOAE) test to diagnose the hearing impairment. The audiology data is reported by stover *et al.* [42] and Dodd and Pepe [38]. The real dataset includes 489 hearing impaired and 1359 normally hearing subjects who were examined at three frequency (f) and three intensity (L) settings of the DPOAE device. Each subject was tested in only one ear. The test result is determined by the negative signal to noise ratio, -SNR. In this real example, the covariates are selected to be $X_f = frequency\ HZ/100$, $X_L = intensity\ dB/10$, $X_D = (hearing\ threshold - 20)dB/10$. If the audiometric threshold is greater than 20 dB HL, the disease status variable $D = 1$; otherwise $D = 0$.

Then the AUC regression model is based on the logarithm odds function:

$$\log(\frac{AUC}{1 - AUC}) = \beta_0 + \beta_1 * X_D + \beta_2 * X_L + \beta_3 * X_f.$$

By the usual GLM-based estimation method, we obtain the estimate for the parameter vector in the AUC regression as shown in Table 3.5.

Based on the estimates for the parameters obtained in the AUC regression, we can find that (1) the AUC odds increase by 3.29% for every 10 dB increase in the hearing threshold

Table (3.5) Real example: the parameters estimates in the AUC regression

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
2.376277	0.03239957	-0.04215329	0.000273678

(AUC odds, 1.0329); (2) the AUC odds decrease by 4.13% for every 10 dB increase in intensity (AUC odds, 0.9587); (3) the AUC odds increase 0.03% for every 100 HZ increase in the frequency. These results are coincide with the result of Dodd and Pepe [38].

Applying the proposed BIFEL and JEL methods, we can construct 95% confidence regions for the parameters vector β in the AUC regression as follows:

95% BIFEL region: $CR(\beta) = \{\beta : l(\beta) \leq 1741.779\}$,

95% JEL region: $CR_J(\beta) = \{\beta : l_J(\beta) \leq 4080.825\}$,

In order to construct confidence intervals for the AUC at a specified value of the covariate vector $\mathbf{Z_D} = (\mathbf{X_D}, \mathbf{X_L}, \mathbf{X_f})$, we choose covariate vector $\mathbf{Z_D}$ to be the median of the observed values for the covariate vector $\mathbf{Z_D}$, and choose $B = 1000$. Then applying the procedure introduced in section 3.6 and the method proposed in sections 3.4 & 3.5, we obtain that the estimate for the covariate-specific AUC is 0.9271 at the median covariates. The 95% BIFEL interval for the AUC is $(0.8624, 1)$, and the 95% JEL interval for the AUC is $(0.8227, 1)$. Hence if the covariate vector $\mathbf{Z_D}$ is set to be the median value of the covariates, the test has relatively moderate to high diagnostic accuracy.

PART 4

EMPIRICAL LIKELIHOOD-BASED INTERVAL ESTIMATION FOR THE CORRELATION COEFFICIENT

4.1 Introduction

In literature, many researchers have focused on developing the theoretical methodology for the correlation between variables. Sir Francis Galton [43] pioneered the theoretical concept of bivariate correlation. Pearson [44] defined a product-moment correlation coefficient which is an index still in use for quantifying the association between two variables. Pearson [45] published a paper entitled "Notes on the History of Correlation" and credited Carl Friedrich Gauss for developing the normal surface of n correlated variates. Correlation is a statistic that measures the degree to which two variables move in relation to each other. One important index related is correlation coefficient (CC), which is a commonly used measure of a possible linear relationship between two continuous random variables. It is a very popular tool for analyzing data that arise in many scientific disciplines such as biology, biomedical and medical research, economics, and agriculture. For more details on the history of the correlation coefficient, we refer readers to Rogers and Nicewander [46].

4.2 The Goal of this Part

Confidence intervals for the CC can be constructed when the underlying distribution is a bivariate normal distribution. However, if the joint distribution of two variables is not normal or unknown, parametric inferences on ρ become quite difficult. On the other hand, non-parametric inferences on the correlation do not need the assumption of bivariate normality for the underlying distribution of (X, Y). Therefore, the primary goal of this part is to propose new non-parametric confidence intervals for the CC.

In more details, it is shown that the asymptotic distribution of the influence function-

based empirical log-likelihood ratio statistic is a standard chi-square distribution. Hence, confidence intervals can be easily obtained without any complicated density function estimates. So we focus on constructing the influence function-based confident interval for CC. In the methodology part, we will propose two EL-based intervals for the CC, including a plug-in EL-based interval and an influence function-based EL Interval. Then extensive simulation studies are conducted to examine the finite sample performances of the proposed intervals compared with the existing parametric and non-parametric intervals for the CC. Afterwards, two real examples are used to illustrate our proposed methods. Finally, we conclude this part with brief discussion. The proof of the main theorem is deferred until Appendix.

4.3 Methodology

Let (X, Y) be a bivariate random vector with mean μ and covariance matrix Σ:

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}.$$

The correlation coefficient ρ between X and Y is defined as follows:

$$\rho = \frac{E(X-\mu_x)(Y-\mu_y)}{\sqrt{E(X-\mu_x)^2}\sqrt{E(Y-\mu_y)^2}} \equiv \frac{\sigma_{xy}}{\sigma_x \sigma_y}. \tag{4.1}$$

Assume that (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d. observations for (X, Y). Pearson's [44] product-moment correlation coefficient is

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}, \tag{4.2}$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. r is a consistent estimate for the CC ρ. It measures the linear association between X_i's and Y_i's.

When (X, Y) follows a bivariate normal distribution, Fisher [47] expressed the density

function of r as

$$f_1(r) = \frac{2^{n-3}(1-\rho^2)^{\frac{n-4}{2}}(1-r^2)^{\frac{n-4}{2}}}{(n-3)!\pi}\sum_{j=0}^{\infty}\frac{(2\rho r)^j}{j!}\Gamma^2[\frac{1}{2}(n-1+j)]. \tag{4.3}$$

Hotelling [48] modified (4.3) and obtained a new expression for the density of r:

$$f_2(r) = \frac{n-2}{\sqrt{2\pi}}\frac{\Gamma(n-1)}{\Gamma(n-\frac{1}{2})}(1-\rho^2)^{\frac{n-4}{2}}(1-r^2)^{\frac{n-4}{2}}(1-\rho r)^{-n+\frac{3}{2}}G(\frac{1}{2},\frac{1}{2};n-\frac{1}{2},\frac{1+\rho r}{2}), \tag{4.4}$$

where $G(a,b;c,x) = \sum_{j=0}^{\infty}\frac{\Gamma(a+j)}{\Gamma(a)}\frac{\Gamma(b+j)}{\Gamma(b)}\frac{\Gamma(c)}{\Gamma(c+j)}\frac{x^j}{j!}$ is the Gaussian hypergeometric function. Since $G(a,b;c,x)$ converges more quickly than the infinite series in (4.3), the function in (4.4) provides a better approximation to the density function with smaller n. Confidence intervals for the CC may be developed by using these density functions. However, finding confidence intervals for the CC is computationally cumbersome based on either of these functions.

4.3.1 Plug-in Empirical Likelihood-based Interval for CC

Empirical likelihood (EL), has been used heuristically for purposes of non-parametric estimation of parameters of interest. Owen [12][13] showed that EL ratio statistics for various parameters of an unknown distribution have certain chi-square distributions and may be used to obtain confidence intervals in a way that is completely analogous to that used with parametric likelihoods. We find that EL for parameters can be developed and shown to have properties similar to those for parametric likelihood which has robust property against the underlying distribution function. The advantages of the EL-based methods are summarized as follows: (1) EL-based intervals do not require a pivotal statistic; (2) No prior constraints for the shape of confidence intervals are needed; (3) EL-based intervals are associated with a Bartlett correction that tolerates low coverage error (See Hall and La Scala, [14]). By these advantages, EL method has been widely applied to various fields of scientific research. As we all know, in literature many methods have been proposed for constructing confidence intervals of the CC, such as Z-transformation based confident interval, maximum likelihood

based confident interval. However, EL-based methods for the CC have not been developed well. So, in this section, we attempt to apply the EL method to the construction of confidence intervals for the CC. Note that

$$\rho = E(\frac{X-\mu_x}{\sigma_x} \cdot \frac{Y-\mu_y}{\sigma_y}) = \frac{E(XY) - E(X)E(Y)}{\sqrt{[E(X^2) - E(X)^2][E(Y^2) - E(Y)^2]}}$$

is a smooth function of the mean vector $\mathbf{m} = (E(X), E(Y), E(X^2), E(Y^2), E(XY))$. It can be shown that the asymptotic distribution of the empirical log-likelihood ratio for $\mathbf{m}$ is a chi-square distribution with 5 degrees of freedom. Therefore, one can use this chi-square distribution to obtain an EL-based confidence region for $\mathbf{m}$, and then find an EL-based confidence interval for the CC. However, this method involves complicated computation of a confidence region with five dimensions. Here we propose a plug-in EL confidence interval for the CC which can be easily implemented in practice.

Let $W_i = (X_i, Y_i),\ i = 1, \ldots, n$. Since the CC ρ satisfies the following equation:

$$E\left(\frac{X-\mu_x}{\sigma_x} \cdot \frac{Y-\mu_y}{\sigma_y} - \rho\right) = 0,$$

then the EL for ρ can be defined as follows:

$$L_0(\rho) = \sup_{\mathbf{p}} \left\{ \prod_{i=1}^{n} p_i :\ p_i \geq 0,\ \sum_{i=1}^{n} p_i = 1,\ \sum_{i=1}^{n} p_i (V(W_i) - \rho) = 0 \right\}, \tag{4.5}$$

where $\mathbf{p} = (p_1, \ldots, p_n)$ is a probability vector, and $V(W_i) = \frac{X_i - \mu_x}{\sigma_x} \cdot \frac{Y_i - \mu_y}{\sigma_y},\ i = 1, \ldots, n$.

The population means (μ_x, μ_y) and the population standard deviations (σ_x, σ_y) are unknown in practice, but (μ_x, μ_y) can be estimated by the sample means $(\bar{X}, \bar{Y})$, and (σ_x, σ_y) can be estimated by the sample standard deviations (S_X, S_Y). After plugging these estimates in (4.5), we get the following plug-in EL for the CC:

$$\hat{L}(\rho) = \sup_{\mathbf{p}} \left\{ \prod_{i=1}^{n} p_i :\ p_i \geq 0,\ \sum_{i=1}^{n} p_i = 1,\ \sum_{i=1}^{n} p_i (\hat{V}(W_i) - \rho) = 0 \right\}$$

where $\hat{V}(W_i) = \frac{X_i - \bar{X}}{S_X} \cdot \frac{Y_i - \bar{Y}}{S_Y},\ i = 1, \ldots, n.$

We can simply obtain the following expression for p_i by using the Lagrange multiplier method:

$$p_i = \frac{1}{n}\{1 + \lambda(\hat{V}(W_i) - \rho)\}^{-1}, \quad i = 1, \ldots, n,$$

where λ is a solution of the following equation:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{V}(W_i) - \rho}{1 + \lambda(\hat{V}(W_i) - \rho)} = 0. \tag{4.6}$$

Subject to $\sum_{i=1}^{n} p_i = 1$, $\prod_{i=1}^{n} p_i$ has its maximum n^{-n} at $p_i = n^{-1}$. Hence, the plug-in EL ratio for ρ has the following expression:

$$R(\rho) = \prod_{i=1}^{n} (np_i) = \prod_{i=1}^{n} \{1 + \lambda(\hat{V}(W_i) - \rho)\}^{-1}. \tag{4.7}$$

The corresponding plug-in empirical log-likelihood ratio for ρ is given by

$$\ell(\rho) = -2 \log R(\rho) = 2 \sum_{i=1}^{n} \log\{1 + \lambda(\hat{V}(W_i) - \rho)\}. \tag{4.8}$$

Theorem 4.1. If ρ is the true value of the correlation coefficient, then the asymptotic distribution of $\ell(\rho)$ is a scaled chi-square distribution with one degree of freedom. i.e.,

$$A \cdot \ell(\rho) \xrightarrow{d} \chi_1^2,$$

where the scale constant $A = \sigma_0^2/\sigma_V^2$ with

$$\begin{aligned} \sigma_V^2 &= Var[X - \mu_x\sigma_x \cdot \frac{Y - \mu_y}{\sigma_y} - 2^{-1}\rho((X - \mu_x\sigma_x)^2 + (Y - \mu_y\sigma_y)^2)], \\ \sigma_0^2 &= Var[X - \mu_x\sigma_x \cdot \frac{Y - \mu_y}{\sigma_y}]. \end{aligned}$$

In order to construct a confidence interval for ρ based on Theorem 1, we need to estimate

the scale constant A. Let

$$\begin{aligned} A_{1i} &= \frac{X_i-\bar{X}}{S_X}\cdot\frac{Y_i-\bar{Y}}{S_Y} - 2^{-1}r((\frac{X_i-\bar{X}}{S_X})^2+(\frac{Y_i-\bar{Y}}{S_Y})^2), \\ A_{2i} &= \frac{X_i-\bar{X}}{S_X}\cdot\frac{Y_i-\bar{Y}}{S_Y}, \\ \widehat{\sigma}_V^2 &= \frac{1}{n}\sum_{i=1}^{n}(A_{1i}-\frac{1}{n}\sum_{i=1}^{n}A_{1i})^2, \\ \widehat{\sigma}_0^2 &= \frac{1}{n}\sum_{i=1}^{n}(A_{2i}-\frac{1}{n}\sum_{i=1}^{n}A_{2i})^2. \end{aligned}$$

Then, $\widehat{A}=\widehat{\sigma}_0^2/\widehat{\sigma}_V^2$ is a consistent estimate for the scale constant A, and a $(1-\alpha)$ level plug-in EL-based confidence interval (called PEL interval) for ρ can be constructed as follows:

$$\{\rho : \widehat{A}\cdot\ell(\rho)\le\chi_1^2(1-\alpha)\},$$

where $\chi_1^2(1-\alpha)$ is the $(1-\alpha)$-th quantile of χ_1^2.

We need the following lemma for the proof of Theorem 4.1.

Lemma 4.1.

(i). $n^{-1/2}\sum_{i=1}^{n}\widehat{V}(W_i)-\rho \xrightarrow{\mathcal{L}} N(0,\sigma_V^2)$,

where $\sigma_V^2 = Var[\frac{X-\mu_x}{\sigma_x}\cdot\frac{Y_i-\mu_y}{\sigma_y} - 2^{-1}\rho((\frac{X-\mu_x}{\sigma_x})^2+(\frac{Y-\mu_y}{\sigma_y})^2)]$.

(ii). $\frac{1}{n}\sum_{i=1}^{n}(\widehat{V}(W_i)-\rho)^2 \xrightarrow{p} \sigma_0^2$, where $\sigma_0^2 = Var[\frac{X-\mu_x}{\sigma_x}\cdot\frac{Y_i-\mu_y}{\sigma_y}]$.

The proof of Lemma 4.1 and Theorem 4.1 are deferred until Appendix.

4.3.2 Influence Function-based Empirical Likelihood Interval for CC

Theorem 4.1 tells us empirical log-likelihood ratio statistic for CC is a scaled chi-square distribution and can be used to obtain confidence interval in a way that is completely analogous to that used with parametric likelihood. In this section, we will define a new function called influence function, so that the asymptotic distribution of the empirical log-likelihood ratio of parameter of interest CC is just a standard chi-squared distribution with one degree

of freedom.

From the proof of Lemma 4.1 showed in the appendix, we get the following expression:

$$n^{-1/2}\sum_{i=1}^{n}(V(W_i)-\rho) \;\; = \;\; n^{-1/2}\sum_{i=1}^{n}V_I(W_i,\rho)+o_p(1)\xrightarrow{\mathcal{L}} N(0,\sigma_V^2),$$

where

$$V_I(W_i,\rho)=(\frac{X_i-\mu_x}{\sigma_x}\cdot\frac{Y_i-\mu_y}{\sigma_y}-\rho)-2^{-1}\rho[((\frac{X_i-\mu_x}{\sigma_x})^2-1)+((\frac{Y_i-\mu_y}{\sigma_y})^2-1)]$$

is the influence function for ρ.

Let $\mathbf{p}=(p_1,\ldots,p_n)$ be a probability vector. Based on this influence function for ρ, we can define an influence function-based EL for ρ as follows:

$$L_I(\rho)=\sup_{\mathbf{p}}\left\{\prod_{i=1}^{n}p_i:\ p_i\geq 0,\ \sum_{i=1}^{n}p_i=1,\ \sum_{i=1}^{n}p_i\widehat{V}_I(W_i,\rho)=0\right\},$$

where

$$\widehat{V}_I(W_i,\rho)=(\frac{X_i-\bar{X}}{S_X}\cdot\frac{Y_i-\bar{Y}}{S_Y}-\rho)-2^{-1}\rho[((\frac{X_i-\bar{X}}{S_X})^2-1)+((\frac{Y_i-\bar{Y}}{S_Y})^2-1)].$$

We can simply obtain the following expression for p_i by using the Lagrange multiplier method:

$$p_i=\frac{1}{n}\{1+\lambda_I V_I(W_i,\rho)\}^{-1},\quad i=1,\ldots,n,$$

where λ_I is a solution of the following equation:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{V}_I(W_i,\rho)}{1+\lambda_I\widehat{V}_I(W_i,\rho)}=0. \tag{4.9}$$

The corresponding influence function-based empirical log-likelihood ratio for ρ is given by

$$\ell_I(\rho)=2\sum_{i=1}^{n}\log\{1+\lambda_I\widehat{V}_I(W_i,\rho)\}. \tag{4.10}$$

Theorem 4.2. If ρ is the true value of the correlation coefficient, then the asymptotic distribution of $\ell_I(\rho)$ is a chi-square distribution with one degree of freedom. i.e.,

$$\ell_I(\rho) \xrightarrow{d} \chi_1^2.$$

Using Theorem 4.2, a $(1-\alpha)$ level influence function-based empirical likelihood confidence interval (called IFEL interval) for ρ can be constructed as follows:

$$\{\rho : \ell_I(\rho) \leq \chi_1^2(1-\alpha)\}.$$

4.4 Simulation Studies

In this section, simulation studies are conducted to examine the finite sample performances of the proposed influence function EL-based confidence intervals. We compared the new confident interval with existing confident intervals. These intervals are namely the normal approximation (NAI) interval based on Fisher's Z-transformation, the hybrid maximum likelihood and bootstrap interval (NAII), the Generalized Pivotal Quantity (GPQ) based interval, the plug-in EL-based interval (PEL), and the influence function EL-based interval (IFEL).

The most commonly used method for the calculation of a confidence interval for the CC is Z-transformation-based method. This method is proposed by Fisher [49] in 1921. However, in 1953 Hotelling found that Fisher's Z-transformation is prone to significant errors when sample size becomes small, so he derived a modification of Fisher's Z-transformation. Weerakkody and Givaruangsawat [50] estimated the CC in the presence of correlated observations from a bivariate normal population. Sun and Wong [51] developed a likelihood-based higher-order asymptotic method to obtain confidence intervals for the correlation coefficient. They recommended the modified signed log-likelihood ratio method over the other approximated methods. They found that approximated methods by Fisher [47] and Hotelling [48] gave very good coverage probabilities but had asymmetric error probabilities, and the ap-

proximated method by Ruben (1966) was not satisfactory. Recently, Tian and Wilding [22] showed that generalized pivotal quantity developed by Weerahandi [52] can be applied towards constructing confidence intervals for the CC.

4.4.1 Z-transformation-based Confidence Intervals

Z-transformation based confident intervals have been introduced by many authors. Fisher [47] and Hotelling [48] derived exact forms of the density function of the sample correlation coefficient. However, obtaining the confidence intervals for the CC based on each form of the density function is computational intensive. Nie *et al.* [53] showed that the maximum likelihood estimator (MLE) r of the correlation coefficient ρ is asymptotically normal with variance $var(r) = (1-\rho^2)^2$. Then $\sqrt{n}(r-\rho)/\sqrt{var(r)}$ asymptotically follows $N(0,1)$ as $n \to \infty$. The density function of r is highly skewed as $|\rho|$ is close to 1. To reduce the skewness of the distribution of r, Fisher's [49] Z-transformation, which is defined by $Z(r) = tanh^{-1}(r) = \frac{1}{2}log\frac{1+r}{1-r}$, can be used to construct a confidence interval for CC ρ.

Since $Z(r)$ is asymptotically a normal distribution with mean $\zeta = \frac{1}{2}log\frac{1+\rho}{1-\rho}$ and variance $\frac{1}{n-3}$, then we have that

$$Z \;=\; \frac{\sqrt{n-3}}{2}\log\frac{(1+r)(1-\rho)}{(1-r)(1+\rho)} \xrightarrow{d} N(0,1).$$

Therefore, we can construct a $(1-\alpha)$ level Z-transformation-based confidence interval (called NAI interval) for the CC as follows:

$$I = \left\{\rho : \frac{B_1 - 1}{B_1 + 1} \le \rho \le \frac{B_2 - 1}{B_2 + 1}\right\},$$

where $B_1 = \frac{1-r}{1+r}e^{-\frac{2}{\sqrt{n-3}}Z_{\alpha/2}}$, $B_2 = \frac{1-r}{1+r}e^{\frac{2}{\sqrt{n-3}}Z_{\alpha/2}}$, $Z_{\alpha/2}$ is $(1-\alpha/2)$-th quantile of the standard normal distribution. This confidence interval has good performance for the relatively large sample sizes, particularly for $n \ge 30$, but it doesn't work very well with small sample sizes.

4.4.2 Maximum Likelihood-based Confidence Intervals

We introduced Z-transformation based methods in the last section but it can only be applied towards the construction of confident interval for CC when underlying distribution is bivariate normal. However, the normality assumption can not be guaranteed in practice. So in this section, we will apply another normal approximation-based method, the maximum likelihood method (MLE), to develop a confidence interval for the CC, assuming that the underlying distribution belongs to a specific bivariate parametric family.

In the following, we choose the bivariate exponential distribution as an example. Bivariate exponential distribution is an important and popular parametric distribution in reliability theory and survival analysis. Different forms of bivariate exponential distributions exist in literature such as those of Gumbel (1960), Freund (1961), Marshall and Olkin (1967) and Block and Basu (1974). Also researchers proposed various forms of both bivariate and multivariate exponential distributions and provided many of their useful properties (Freud, [54]). The most commonly used density function of a bivariate exponential distribution is proposed by Downton [55]. He provided the following specific form of probability density function for bivariate exponential distribution:

$$f(x_1, x_2; \theta) = \frac{\mu_1 \mu_2}{1-\rho} e^{-\frac{\mu_1 x_1 + \mu_2 x_2}{1-\rho}} I_0 \frac{2(\rho \mu_1 \mu_2 x_1 x_1 \rho)^{1/2}}{1-\rho}, \tag{4.11}$$

where $\theta = (\mu_1, \mu_2, \rho)$, $\mu_1, \mu_2 > 0, 0 \leq \rho < 1$, and $I_0(z) = \sum_{r=0}^{\infty} (\frac{z}{2})^{2r}/r!^2$ is the modified Bessel function. The parameter ρ in (4.11) measures the correlation between two bivariate exponential variables.

Let (X_i, Y_i), $i = 1, \ldots, n$ be a random sample drawn from Downton's bivariate exponential distribution. Then the likelihood function of θ based on (X_i, Y_i)'s is

$$L(\theta) = \prod_{i=1}^{n} f(X_i, Y_i; \theta).$$

Al-ssadi and Young [56] derived the maximum likelihood estimator (MLE) $\hat{\theta}$ for θ by

maximizing $L(\theta)$. They showed that $\hat{\theta} = (\hat{\mu}_1, \hat{\mu}_2, \hat{\rho})^T$, and proved that $(\hat{\mu}_1 - \mu_1, \hat{\mu}_2 - \mu_2, \hat{\rho} - \rho)^T$ asymptotically follows the normal distribution $N_3(0, V^{-1})$ where $V = (V_{i,j})$ is a 3×3 matrix with

$$\begin{aligned}
V_{1,1} &= \frac{n}{\mu_1^2}\frac{1-2\rho-\rho^2}{(1-rho)^2} + \frac{\rho^2 A(\rho)}{(1-\rho)^4}, \quad V_{1,2} = \frac{n}{\mu_1\mu_2}\frac{\rho^2 A(\rho)}{(1-\rho)^4} + \frac{\rho(1+\rho)}{(1-\rho)^2}, \\
V_{1,3} &= \frac{n}{\mu_1}\frac{\rho(1+\rho)A(\rho)}{(1-\rho)^5} + \frac{\rho(3+\rho)}{(1-\rho)^3}, \; V_{2,2} = \frac{n}{\mu_2^2}\frac{1-2\rho-\rho^2}{(1-\rho)^2} + \frac{\rho^2 A(\rho)}{(1-\rho)^4}, \\
V_{2,3} &= \frac{n}{\mu_2}\frac{\rho(1+\rho)A(\rho)}{(1-\rho)^5} + \frac{\rho(3+\rho)}{(1-\rho)^3}, \; V_{3,3} = n\frac{(1+\rho)^2 A(\rho)}{(1-\rho)^6} - \frac{(1+\rho)(3+\rho)}{(1-\rho)^4},
\end{aligned}$$

where $A(\rho) = (1-\rho)^5 \int_0^\infty \int_0^\infty \frac{g'^2(\rho y_1 y_2)}{g(\rho y_1 y_2)} y_1^2 y_2^2 e^{-(y_1+y_2)} dy_1 dy_2$, and $g(z) = I_0(2z^{1/2})$.

We can see that the asymptotic variance of $\hat{\theta}$ is of a very complex form. Although one can apply delta method to construct a confidence interval for ρ using above asymptotic normal distribution of $\hat{\theta}$, the method requires intensive computation and plug-in estimation of unknown parameters, and the resulting asymptotic variance estimate is unstable, particularly when sample size n is small. Therefore, we recommend using the bootstrap procedure to estimate the asymptotic variance of $\hat{\rho}$. By generating B (We recommend $B \geq 200$) times of replications $\hat{\rho}_b^*$ of $\hat{\rho}$, for $b = 1 \ldots B$, we can define the bootstrap estimate for the asymptotic variance of $\hat{\rho}$ as

$$Var(\hat{\rho}^*) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\rho}_b^* - \frac{1}{B}\sum_{b=1}^{B}\hat{\rho}_b^*\right)^2.$$

Then a $(1-\alpha)$ level normal approximation-based confidence interval (called NAII) for the CC can be constructed as follows, which is a hybrid maximum likelihood and bootstrap interval:

$$(\hat{\rho} - Z_{\alpha/2}\sqrt{Var(\hat{\rho}^*)}, \hat{\rho} + Z_{\alpha/2}\sqrt{Var(\hat{\rho}^*)}),$$

where $Z_{\alpha/2}$ is $(1-\alpha/2)$-th quantile of the standard normal distribution.

4.4.3 Generalized Confidence Interval

Let X be an observable random vector with the cdf $F(x|v)$, where $v = (\theta, \delta)$ is a vector of unknown parameters, θ is the parameter of interest, and δ is a vector of nuisance parameters. Let χ be the sample space of possible values of X and let Θ be the parameter space of θ. An observation from X is denoted by x, where $x \in \chi$. Let $R = r(X; x, v)$ be a function of X, x and v. R is said to be a generalized pivotal quantity if R has a probability distribution free of unknown parameters, and the observed pivotal, defined as $r_{obs} = r(X; x, v)$, does not depend on the nuisance parameter δ .

Then a two-sided $100(1-\alpha)\%$ GPQ-based confidence interval for the parameter θ is $(R_{\alpha/2}, R_{1-\alpha/2})$, where $R_{\alpha/2}$ and $R_{1-\alpha/2}$ are the $100(\alpha/2)$-th percentile and $100(1-\alpha/2)$-th percentile of the distribution of R, respectively. More detailed introduction of GPQ-based confidence intervals can be found in Weerahandi [52] and Hanning et al. [57].

Let $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ be a random sample from a bivariate normal distribution with mean μ and covariance matrix Σ. An estimator for (μ, Σ) is

$$(\hat{\mu}, \hat{\Sigma}) = \left(\begin{pmatrix} \overline{X} \\ \overline{Y} \end{pmatrix}, \frac{1}{n-1} \begin{pmatrix} SSx & SSxy \\ SSxy & SSy \end{pmatrix} \right), \tag{4.12}$$

where SSx, SSy and $SSxy$ are sum of squares of X, Y and XY, respectively.

While the CC ρ is a function of parameters $(\sigma_x, \sigma_y, \sigma_{xy})$ under normality assumption. The function of parameters of interest are defined as: $(\mu, \beta, \sigma^2_{2|1}) \equiv (\mu, \frac{\sigma_{xy}}{\sigma^2_x}, \sigma^2_y - \frac{\sigma^2_{xy}}{\sigma^2_x})$. Where the parameters can be estimated by $(Mu, Beta, SSx_{2|1}) \equiv (\hat{\mu}, \frac{SSxy}{SSx}, SSy - \frac{SSxy^2}{SSx})$.

Then we can define the following pivotal quantities U, $U_{2|1}$ and V_B based on the statistics

Mu, $Beta$, SSx, SSy and $SSx_{2|1}$:

$$U = \frac{SSx}{\sigma_x^2} \sim \chi^2_{n-1}, \tag{4.13}$$

$$U_{2|1} = \frac{SSx_{2|1}}{\sigma^2_{2|1}} \sim \chi^2_{n-2}, \tag{4.14}$$

$$V_B = (Beta - \beta)\sqrt{\frac{SSx}{\sigma^2_{2|1}}} \sim N(0,1). \tag{4.15}$$

Let mu, $beta$, $ssx_{2|1}$, ssx and ssy be the observed values of Mu, $Beta$, $SSx_{2|1}$, SSx and SSy, respectively. Then the GPQ for $(\sigma_x^2, \sigma^2_{2|1}, \beta)$ is

$$(R_{\sigma_x^2}, R_{\sigma^2_{2|1}}, R_\beta) = \left(\frac{ssx}{U}, \frac{ssx_{2|1}}{U_{2|1}}, beta - V_b\sqrt{\frac{1}{U_{2|1}}\frac{ssx_{2|1}}{ssx}}\right), \tag{4.16}$$

Therefore the GPQs for $(\sigma_y^2, \sigma_{xy})$, Σ and CC ρ are:

$$(R_{\sigma_y^2}, R_{\sigma_{xy}}) = (R_\beta^2 R_{\sigma_x^2} + R_{\sigma^2_{2|1}}, R_\beta R_{\sigma_x^2}), \tag{4.17}$$

$$R_\Sigma = \begin{pmatrix} R_{\sigma_x^2} & R_{\sigma_{xy}} \\ R_{\sigma_{xy}} & R_{\sigma_y^2} \end{pmatrix},$$

$$R_\rho = \frac{R_{\sigma_{xy}}}{\sqrt{R_{\sigma_x^2} R_{\sigma_y^2}}}. \tag{4.18}$$

The following Monte-Carlo algorithm procedures can be applied to generate values of R_ρ:

1. Calculate the sample mean and covariance matrix from the original sample (X_i, Y_i)'s.
2. Generate U, $U_{2|1}$ and V_B using (4.13)-(4.15).
3. Calculate R_ρ using (4.16)-(4.18).
4. Repeat the above 1 - 3 for $K = 10,000$ times to obtain K values of R_ρ.

Then we can estimate the distribution of R_ρ based on the $K = 10,000$ generated values of R_ρ. A $(1-\alpha)$ level GPQ-based confidence interval for the CC can be constructed as $\left(R_{\alpha/2},\ R_{1-\alpha/2}\right)$, where $R_{\alpha/2}$ and $R_{1-\alpha/2}$ are the $(\alpha/2)$-th and $(1-\alpha/2)$-th quantiles of the $10,000$ values of R_ρ, respectively.

In the simulation studies, we select the sample size $n = 30,\ 50,\ 100$. We set the CC ρ to be 0.1, 0.5 and 0.9 to investigate the performance of the proposed methods, when the dependence between X and Y is weak, moderate, and strong, respectively. We compare the coverage probabilities and average confident lengths by the proposed methods with the existing methods, including NAI, NAII, GPQ, PEL and IFEL confident intervals, under the following four different scenarios. In the studies, we consider the following underlying distributions, where three scenarios are under bivariate normal distributions and one scenario is under bivariate exponential distribution.

I. Bivariate Normal Distributions:

- Scenario 1: $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{xy} \\ \sigma_{xy} & 1 \end{pmatrix} \right)$

- Scenario 2: $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{xy} \\ \sigma_{xy} & 2 \end{pmatrix} \right)$

- Scenario 3: $\begin{pmatrix} X \\ Y \end{pmatrix} \sim 0.9{*}N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{xy} \\ \sigma_{xy} & 1 \end{pmatrix} \right) + 0.1{*}N_2 \left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{xy} \\ \sigma_{xy} & 1 \end{pmatrix} \right)$, which is a mixed normal distribution with 90% of observations from the first normal distribution and 10% of observations from the second normal distribution.

II. Bivariate Exponential Distribution:

- Scenario 4 : $(X, Y) \sim Biexp\ (\lambda_1, \lambda_2, \lambda_{12})$, where $\lambda_1, \lambda_2, \lambda_{12}$ are parameters of the bivariate exponential distribution used in Marshall *et al.* [58].

Simulated samples are generated in three scenarios to examine the performance of the confidence intervals of CC ρ under bivariate normal distributions and one scenario with a bivariate exponential distribution. Particularly, scenario 1 has bivariate normal distribution of the same mean, and same variance; scenario 2 has different means, and different variances; and scenario 3 has a normal mixture model with different means and same variance since sometimes the underlying distribution is miss-specified. In scenario 4, we choose a bivariate exponential distribution to generate random samples and evaluate the proposed methods when the underlying distribution is not normal. The following algorithm procedures proposed by Marshall *et al.* [58] are used to generate bivariate exponential random variates with a specific correlation.

(1) Generate random samples $U_1 \sim exp(\lambda_1), U_2 \sim exp(\lambda_2)$, and $U_3 \sim exp(\lambda_{12})$, where

$$\rho = \frac{\lambda_{12}}{(\lambda_1 + \lambda_2 + \lambda_{12})}.$$

(2) Set $X = min(U_1, U_3)$, $Y = min(U_2, U_3)$.

(X_i, Y_i)'s generated by using this algorithm is a random sample from the bivariate exponential distribution with CC ρ.

For each scenario, we run the simulation for 10,000 times to calculate the coverage probability and average length of the above mentioned intervals. B= 500 bootstrap replications are used for the calculation of the NAII intervals. The simulation results are shown in Tables 4.1–4.4. From Tables 4.1–4.3, we observe that among the three parametric intervals, NAI and GPQ intervals have similar performances in most cases. The coverage probabilities of NAI and GPQ intervals are very close to the nominal level 0.95, except for the cases with $\rho = 0.9$ under scenario 3 in which NAI and GPQ intervals have severe under-coverage problems, which indicates they are not robust to the model miss-specification. NAI and GPQ intervals outperform NAII intervals in most cases except for the cases with $\rho = 0.9$ in scenario 3. Comparing the two non-parametric intervals, IFEL intervals outperform PEL intervals in terms of coverage probability and interval length in most cases under all the scenarios. In

scenarios 1–2, the parametric NAI and GPQ intervals have slightly better coverage accuracy that the non-parametric IFEL intervals. However, the non-parametric IFEL intervals perform much better than the parametric NAI, NAII and GPQ intervals in scenario 3. When the normality assumption is violated under scenario 4, NAI and GPQ intervals cannot be used, the non-parametric IFEL intervals outperform the other intervals. In summary, NAI, GPQ and IFEL intervals are recommended when the underlying distribution is a normal distribution, and IFEL interval is preferred when the underlying distribution is non-normal or unknown.

4.5 Real Data Analysis

In this section, we apply two real examples to demonstrate the application of our proposed methods.

Example 1. Brain Size and Intelligence

Willerman *et al.* [59] did research on the association between brain size and mental capacity. In the study, there are 40 right-handed introductory psychology students at a southern university involved in an experiment, where they received tests in four different areas based on Wechsler's [60] Adult Intelligence Scale-revised. The test score measured full scale IQ score using a Verbal IQ and a Performance IQ score. In order to measure the brain size of the subjects, Magnetic Resonance Imaging (MRI) was used. The MRI scanned 18 horizontal MR images. Then a computer counted all the pixels with non-zero gray scale resulting in an index of brain sizes.

The dataset includes several other variables like gender and body size, but here we only consider an index of brain size as the predictor of intelligence. The data set is available from the Data and Story Library (DASL) at Carnegie Mellon University. The sample correlation coefficient between brain size and full scale IQ is $r = 0.3337$. Since the p-value = 0.9851 by the Shapiro-Wilk multivariate normality test (Shapiro and Wilk, [61]), we accept the normality assumption for the data. We calculate 95% confident interval under NAI, GPQ

and IFEL methods for ρ. The results are displayed in Table 4.5.

From the results shown in Table 4.5, we see that NAI, GPQ, and IFEL intervals have similar lengths. All these intervals indicate that the correlation between brain size and intelligence is moderate. There is no strong linear association between brain size and mental ability.

Example 2. Reading and Mathematical Performance

The relationship between the language proficiency and mathematical performance is discussed a lot in literature. For instance, MacGregor and Price [62] made a short review of the literature on the connection between reading comprehension and mathematics performance. They indicated that the ability to read, interpret and comprehend the working problems significantly affect the result whether they can perform successfully in mathematics.

In this study, we would like to apply a real dataset from the National Center for Education Statistics (NCES) to construct the confidence interval for correlation coefficient of reading capability and mathematical performance. The dataset (Data source: NCES, http://nces.ed.gov) covers the average scores in Reading and Mathematics for 4^{th} and 8^{th} graders in 11 urban areas, including New York City, Boston, Chicago, etc., in 2005. The sample correlation coefficient is $r = 0.9631$, which indicates a strong relationship between the two variables. Based on the Shapiro-Wilk multivariate test for the normality with p-value equal to 0.0266, we can conclude that dataset follows non-normality for the underlying distribution. We conduct simulation studies to compare IFEL interval with PEL interval for ρ. The results are displayed in the Table 4.6. From the results, we can observe that the 95% IFEL interval for ρ is (0.7808, 0.9680) suggesting that reading and math performances are highly correlated as many authors have mentioned throughout the history of education.

4.6 Discussion

The correlation coefficient is a fundamental measurement for the strength of the linear relationship between two random variables. It's commonly used for analyzing the relationship between two random variables in scientific areas. Many parametric inferences have been proposed for the correlation coefficient so far. However, when the underlying distribution is not normal or unknown, the construction of confident intervals for correlation coefficient are not well developed. Therefore, in the perspective of developing non-parametric inferences for CC, we propose two empirical likelihood-based non-parametric intervals for the correlation coefficient including plug-in empirical likelihood confident interval and influence function based confident interval. It has been shown that empirical likelihood ratio statistics for parameters of an unknown distribution have certain chi-square distributions and may be used to obtain confidence intervals in a way that is completely analogous to that used with parametric likelihoods. We evaluate these intervals through extensive simulation studies and compare our proposed methods with the existing methods in terms of coverage probabilities and average confident interval lengths. The simulation results indicate that the GPQ-based interval performs very well when the underlying distribution is normal while the IFEL interval has better overall performances with finite samples when the underlying distribution is non-normal or unknown. We recommend the use of the Z-transformation based NAI interval, the GPQ interval and the IFEL interval when the underlying distribution is a normal distribution, and the use of IFEL interval when the underlying distribution is non-normal or unknown.

Table (4.1) Coverage probabilities and average lengths of various 95% confidence intervals for ρ, scenario 1.

		Coverage Probability			Average Length		
n	Method	0.1	0.5	0.9	0.1	0.5	0.9
30	NAI	0.958	0.953	0.949	0.699	0.546	0.156
	NAII	0.901	0.911	0.915	0.688	0.532	0.148
	GPQ	0.948	0.951	0.950	0.684	0.545	0.159
	PEL	0.910	0.916	0.946	0.497	0.519	0.130
	IFEL	0.963	0.958	0.949	0.435	0.511	0.147
50	NAI	0.950	0.951	0.951	0.542	0.421	0.114
	NAII	0.928	0.904	0.931	0.535	0.409	0.109
	GPQ	0.948	0.948	0.948	0.537	0.419	0.116
	PEL	0.929	0.939	0.952	0.441	0.424	0.103
	IFEL	0.945	0.957	0.957	0.386	0.419	0.113
100	NAI	0.951	0.952	0.952	0.385	0.296	0.078
	NAII	0.936	0.940	0.949	0.382	0.293	0.076
	GPQ	0.947	0.952	0.953	0.384	0.295	0.078
	PEL	0.940	0.955	0.950	0.353	0.313	0.073
	IFEL	0.958	0.954	0.951	0.320	0.308	0.080

Table (4.2) Coverage probabilities and average lengths of various 95% confidence intervals for ρ, scenario 2.

		Coverage Probability			Average Length		
n	Method	0.1	0.5	0.9	0.1	0.5	0.9
30	NAI	0.951	0.948	0.946	0.694	0.549	0.156
	NAII	0.905	0.911	0.903	0.678	0.521	0.147
	GPQ	0.951	0.949	0.950	0.684	0.544	0.158
	PEL	0.909	0.935	0.941	0.489	0.521	0.129
	IFEL	0.946	0.948	0.957	0.440	0.510	0.146
50	NAI	0.949	0.953	0.947	0.542	0.421	0.114
	NAII	0.931	0.934	0.922	0.536	0.413	0.108
	GPQ	0.950	0.950	0.951	0.536	0.420	0.115
	PEL	0.937	0.949	0.951	0.434	0.423	0.102
	IFEL	0.962	0.951	0.954	0.392	0.419	0.114
100	NAI	0.947	0.948	0.949	0.385	0.296	0.078
	NAII	0.923	0.946	0.947	0.381	0.293	0.076
	GPQ	0.950	0.952	0.949	0.384	0.296	0.078
	PEL	0.949	0.952	0.949	0.357	0.315	0.074
	IFEL	0.948	0.954	0.954	0.319	0.308	0.079

Table (4.3) Coverage probabilities and average lengths of various 95% confidence intervals for ρ, scenario 3.

		Coverage Probability			Average Length		
n	Method	0.1	0.5	0.9	0.1	0.5	0.9
30	NAI	0.955	0.954	0.808	0.695	0.567	0.228
	NAII	0.903	0.934	0.918	0.682	0.551	0.214
	GPQ	0.950	0.947	0.760	0.684	0.563	0.234
	PEL	0.905	0.954	0.867	0.484	0.532	0.182
	IFEL	0.955	0.940	0.923	0.433	0.520	0.205
50	NAI	0.948	0.949	0.708	0.542	0.435	0.164
	NAII	0.919	0.934	0.855	0.531	0.425	0.161
	GPQ	0.955	0.945	0.689	0.538	0.432	0.165
	PEL	0.928	0.957	0.919	0.428	0.440	0.145
	IFEL	0.954	0.946	0.937	0.382	0.430	0.160
100	NAI	0.948	0.941	0.537	0.386	0.305	0.109
	NAII	0.938	0.939	0.747	0.382	0.300	0.108
	GPQ	0.946	0.941	0.530	0.384	0.304	0.109
	PEL	0.942	0.943	0.918	0.353	0.324	0.105
	IFEL	0.957	0.959	0.936	0.315	0.317	0.115

Table (4.4) Coverage probabilities and average lengths of various 95% confidence intervals for ρ, scenario 4.

		Coverage Probability			Average Length		
n	Method	0.1	0.5	0.9	0.1	0.5	0.9
30	NAII	0.865	0.818	0.681	0.689	0.665	0.269
	PEL	0.876	0.946	0.935	0.490	0.631	0.287
	IFEL	0.945	0.956	0.948	0.416	0.574	0.249
50	NAII	0.901	0.870	0.742	0.553	0.553	0.254
	PEL	0.940	0.949	0.939	0.448	0.564	0.226
	IFEL	0.955	0.943	0.956	0.380	0.456	0.230
100	NAII	0.917	0.902	0.781	0.406	0.420	0.198
	PEL	0.951	0.940	0.954	0.371	0.454	0.207
	IFEL	0.956	0.953	0.946	0.323	0.405	0.203

Table (4.5) 95% confidence intervals for ρ in example 1

Methods	Confidence Interval	Length
NAI	(0.0157,0.5904)	0.5747
GPQ	(0.0088, 0.5845)	0.5757
IFEL	(0.0088, 0.5697)	0.5609

Table (4.6) 95% confidence intervals for ρ in example 2

Methods	Confidence Interval	Length
PEL	(0.8038, 0.9555)	0.1517
IFEL	(0.7808, 0.9680)	0.1872

PART 5

EMPIRICAL LIKELIHOOD-BASED INTERVAL ESTIMATION FOR THE COEFFICIENT OF VARIATION

5.1 Introduction

The coefficient of variation (CV), as a popular measure for the relative variation of a random variable, is defined as the ratio of the standard deviation to the mean. It is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another. Suppose that $\{X_1, X_2, \cdots, X_n\}$ is a random sample from the population X with the distribution function $F(x)$, where X is a random variable with mean μ and variance σ^2. Then, the population coefficient of variation is defined as $k = \frac{\sigma}{\mu}$. The CV k can be consistently estimated by $K = \frac{S}{\bar{X}}$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ which is the sample mean and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ which is the sample variance. $\bar{X}$ and S^2 are unbiased estimators for μ and σ^2, respectively. We can see that the higher the CV, the greater the level of dispersion of the variable around it's mean, and vice versa.

The CV is an alternative and more informative variation measure besides the more commonly used measurements of variation such as variance (standard deviation), because when comparing the variation across populations, one needs to standardize and take into account different units. If the units of populations are not matching, it is not appropriate to directly compare the variances or standard deviations. Take one simple example, if we want to analyze the variation in both BMI score and blood pressure of patients, directly comparing the variances or standard deviations doesn't make any sense since the scale of data is different. A lot of similar examples can be found in the scientific study, where the data with different measurement scale makes the direct comparison invalid or misleading. Hence, from the perspective of unitlessness, CV is commonly applied and analyzed by researchers and practitioners to make an important decision, particularly in biology, finance, engineering,

and agricultural fields (Taye and Njuho, [63]).

Although the CV is a popular measurement index for the relative variation of a random variable, there is a lack of work for making inferences on the CV in literature. A possible reason is that the distribution of the sample coefficient of variation K is unknown when the underlying distribution $F(x)$ is unknown. Even when $F(x)$ is a normal distribution, the sampling distribution of K has a complicated form (See Hendricks and Robey, [64]):

$$dF_K(v) = \frac{2}{\pi^{\frac{1}{2}}\Gamma(\frac{n-1}{2})} e^{-\frac{n}{2}\frac{v^2}{\sigma^2}\frac{v^2}{1+v^2}} \frac{v^{n-2}}{(1+v^2)^{\frac{n}{2}}} \sum_{i=0}^{n-1} \frac{(n-1)!\Gamma(\frac{n-i}{2})}{(n-1-i)!i!} \frac{n^{\frac{i}{2}}}{2^{\frac{i}{2}}(\frac{\sigma}{\mu})^i} \frac{1}{(1+v^2)^{\frac{i}{2}}} dv.$$

Suppose that $\{X_1, X_2, \cdots, X_n\}$ is a random sample from the normal distribution with mean μ and variance σ^2, Lehmann [65] derived the sampling distribution of K as

$$\frac{\bar{X}}{\frac{S}{\sqrt{n}}} \sim NCT_{n-1}(\frac{\mu\sqrt{n}}{\sigma}),$$

where $NCT_{n-1}(\frac{\mu\sqrt{n}}{\sigma})$ is a non-central t-distribution with $n-1$ degrees of freedom depending on the non-centrality parameter $\frac{\mu\sqrt{n}}{\sigma}$. Using this non-central t-distribution, Hayter [66] showed how to construct confidence intervals for the CV of a normal distribution. However, cumbersome calculations still lead to intensive computation. Hayter and Kim [67] considered the problem of testing the equality of two CV's.

There have been many researchers in the literature who proposed some simple approximation methods with acceptable coverage accuracy. More recently, Mahmoudvand and Hassani [68] used approximate methods to construct a confidence interval for a CV for normally distributed data, Panichkitkosolkul [69] developed confidence intervals for the CV in a normal distribution with a known population mean. Verrill [70] discussed confidence intervals for the CV when the underlying distribution is a log-normal distribution. Tian [71] provided a generalized confidence interval for the common coefficient of variation.

It appears to be difficult for constructing an exact confidence interval for the CV due to the complexity of the distribution function of K. Therefore, it's necessary for us to develop

new inferential methods for the CV.

5.2 The Motivation of this Part

When the underlying of population is normally distributed, many methods have been proposed to construct confidence intervals for the CV, such as by Z-transformation method or maximum likelihood estimation (MLE) method. However, the normality assumption for the underlying distribution may not be easily guaranteed, and in many cases, the underlying distribution of the population is not normal or unknown, so parametric inferences on the CV become quite involved. Since non-parametric inference on the CV does not need a parametric assumption on the underlying distribution, developing non-parametric empirical methods for construction of the confidence intervals for the CV become inevitable and important. By using side information provided by the influence function, empirical likelihood methods are proposed for inferences of a CV. It is shown that the asymptotic distribution of the influence function-based empirical log-likelihood ratio statistic is a standard chi-square distribution. Hence, confidence intervals for the CV can be easily obtained without any underlying distribution assumption of sample coefficient of variation.

Therefore, the motivation of this part is to propose empirical likelihood-based non-parametric intervals for the CV when the underlying distribution is unknown. In section 5.4 we also review some of the existing confidence intervals for the CV and then conduct simulation studies to compare the proposed EL-based confidence intervals with existing confidence intervals for the CV in terms of coverage probabilities and average interval lengths including Vangel's method and generalized confidence interval; A real example is applied to demonstrate our proposed method, which followed by the conclusions and discussions made in section 5.6.

5.3 Empirical Likelihood-based Intervals

5.3.1 Plug-in Empirical Likelihood-based Interval for a CV

As we have introduced in the section 5.2, EL is a popular non-parametric method for constructing confidence intervals of parameters of interest. With many advantages, EL method has been widely applied to many fields. The advantages of the EL-based methods are summarized as follows: (1) EL-based intervals do not require a pivotal statistic; (2) No prior constraints for the shape of confidence intervals are needed; (3) EL-based intervals are associated with a Bartlett correction that tolerates low coverage error (See Hall and La Scala, [14]). However, EL-based methods for the CV have not been developed. So, in this section, we attempt to apply the EL method to the construction of confidence intervals for the CV. Observing that the CV k satisfies the following equation:

$$E(\sigma - kX) = 0,$$

then, the EL for k can be defined as follows:

$$L(k) = \sup_{\mathbf{P}}\{\prod_{i=1}^{n} p_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i W_i(k) = 0\},$$

where $\mathbf{P} = (p_1, ..., p_n)$ is a probability vector, and $W_i(k) = \sigma - kX_i,\ i = 1, \cdots, n$.

The standard deviation σ is usually unknown in practice, but we can use the sample standard deviation S to estimate it. Then, a plug-in EL for k can be defined as follows:

$$\hat{L}(k) = \sup_{\mathbf{P}}\{\prod_{i=1}^{n} p_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \hat{W}_i(k) = 0\},$$

where $\hat{W}_i(k) = S - kX_i$.

Using Lagrange multiplier method, we can obtain the expression for p_i, which is

$$p_i = \frac{1}{n}\{1 + \lambda \hat{W}_i(k)\}^{-1}, i = 1, ..., n,$$

where λ is the solution to

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{W}_i(k)}{1+\lambda\hat{W}_i(k)}=0. \tag{5.1}$$

Therefore, the plug-in EL ratio for k is:

$$R(k)=\prod_{i=1}^{n}(np_i)=\prod_{i=1}^{n}\{1+\lambda\hat{W}_i(k)\}^{-1}.$$

Then the corresponding empirical log-likelihood ratio for k is

$$l(k)=2\sum_{i=1}^{n}log\{1+\lambda\hat{W}_i(k)\}. \tag{5.2}$$

Theorem 5.1 below shows that $l(k)$ asymptotically follows a scaled chi-square distribution. The proof of Theorem 5.1 is similar to the proof of Theorem 4.1 in Part 4.

Theorem 5.1: If k is the true value of the coefficient of variation, then the asymptotic distribution of $l(k)$, defined by (5.2), is a scaled chi-square distribution with degree of freedom one. i.e.,

$$c\ l(k)\xrightarrow{d}\chi_1^2,$$

where the scale constant $c=\sigma_2^2/_1^2$ with

$$\sigma_1^2=E[\frac{(X-\mu)^2}{2\sigma}-\frac{\sigma}{\mu}\cdot(X-\mu)-\sigma 2]^2,\quad \sigma_2^2=\frac{\sigma^4}{\mu^2}.$$

Since the scale constant c is still unknown, in order to construct a confidence interval for the CV, we have to estimate c.

Let

$$\widehat{\sigma}_1^2=\frac{1}{n}\sum_{i=1}^{n}[\frac{(X_i-\bar{X})^2}{2S}-\frac{S}{\bar{X}}\cdot(X_i-\bar{X})-\frac{S}{2}]^2,\quad \widehat{\sigma}_2^2=\frac{S^4}{\bar{X}^2}.$$

Then, by Slutski's lemma, $\widehat{c}=\widehat{\sigma}_2^2/\widehat{\sigma}_1^2$ is a consistent estimate for c. Based on Theorem 5.1, we can construct a $100(1-\alpha)\%$ level plug-in EL-based confidence interval (PEL interval)

for k as follows:

$$\{k : \widehat{c}l(k) \leq \chi_1^2(1-\alpha)\},$$

where $\chi_1^2(1-\alpha)$ is the $(1-\alpha)$-th quantile of χ_1^2.

Another way to estimate the scale constant c is to use the following bootstrap procedure:

Step 1: Generate a bootstrap sample $\{X_1^*, \cdots, X_n^*\}$ from the original sample $\{X_1, \cdots, X_n\}$.

Step 2: Find a bootstrap estimate $\widehat{c}_b^*$ of $\widehat{c}$:

$$\widehat{c}_b^* = (\widehat{\sigma}_2^*)^2/(\widehat{\sigma}_1^*)^2,$$

where $\widehat{\sigma}_1^*$ and $\widehat{\sigma}_2^*$ are the bootstrap copies of $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$, respectively.

Step 3: Repeat steps 1-2 B times ($B \geq 200$ is recommended) to obtain B bootstrap copies of $\widehat{c}_b^*$: $\widehat{c}_1^*, \cdots, \widehat{c}_B^*$.

Step 4: Estimate the constant c as follows:

$$c^* = \frac{1}{B}\sum_{b=1}^{B} \widehat{c}_b^*.$$

Based on Theorem 5.1 and the estimated constant c^*, we can construct a $100(1-\alpha)\%$ level bootstrap and EL-based confidence interval (BPEL interval) for k as follows:

$$\{k : c^* l(k) \leq \chi_1^2(1-\alpha)\}.$$

Two methods of estimation for the scale constant c provide similar final performance in terms of coverage probabilities. We choose the second way to estimate the scale constant c in the simulation studies.

5.3.2 Influence Function-based Empirical Likelihood Interval for a CV

From the proof of Lemma 5.1 in appendix, we get that

$$n^{-1/2}\sum_{i=1}^{n}\widehat{W}_i(k) \;=\; n^{-1/2}\sum_{i=1}^{n}U(X_i,k)+o_p(1)\xrightarrow{\mathcal{L}}N(0,{}^{2}_{1}),$$

where $U(X_i,k)=\frac{(X_i-\mu)^2}{2\sigma}-k(X_i-\mu)-\frac{\sigma}{2}$ is the influence function for k.

Let $\mathbf{P}=(p_1,...,p_n)$ be a probability vector. Based on the influence function for k, the EL for k can be defined as follows:

$$L_I(k)=\underset{\mathbf{P}}{sup}\{\prod_{i=1}^{n}p_i:p_i\geq 0,\sum_{i=1}^{n}p_i=1,\sum_{i=1}^{n}p_i\hat{U}_i(k)=0\},$$

where $\hat{U}_i(k)=\frac{(X_i-\bar{X})^2}{2S}-k(X_i-\bar{X})-\frac{S}{2}$.

With Lagrange multiplier method, we can obtain the expression for p_i, which is

$$p_i=\frac{1}{n}\{1+\lambda_I\widehat{U}_i(k)\}^{-1},i=1,...,n,$$

where λ_I is the solution to

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{U}_i(k)}{1+\lambda_I\widehat{U}_i(k)}=0. \tag{5.3}$$

Then, the corresponding empirical log-likelihood ratio for k is

$$l_I(k)=2\sum_{i=1}^{n}log\{1+\lambda_I\widehat{U}_i(k)\}. \tag{5.4}$$

Theorem 5.2 shows that $l_I(k)$ asymptotically follows a standard chi-square distribution.

Theorem 5.2: If k is the true value of the coefficient of variation, then the asymptotic distribution of $l_I(k)$, defined by (5.4), is a chi-square distribution with degree of freedom one. i.e.,

$$l_I(k)\xrightarrow{d}\chi^2_1.$$

Therefore, a $100(1-\alpha)\%$ level influence function-based empirical likelihood (IFEL) confidence interval for k can be constructed as follows:

$$\{k : l_I(k) \leq \chi_1^2(1-\alpha)\}.$$

5.3.3 Jackknife Empirical Likelihood-based Interval for a CV

Jackknife Empirical Likelihood (JEL) method is an approach proposed by Jing, Yuan and Zhou (2009) for the construction of confidence intervals for a parameter of interest. One attractive property of the JEL-based method is that the logarithm of the JEL ratio statistic asymptotically follows a standard Chi-square distribution under some regularity conditions. Thus, constructing a JEL-based confidence interval is simple in calculation, which motivate us to develop a JEL-based confidence interval for the CV in this section.

Let $\widehat{V} = \frac{1}{n}\sum_{i=1}^{n}\hat{W}_i(k)$, and $\widehat{V}_i = \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\hat{W}_{j,-i}(k),\ i = 1, \cdots, n$, where $\hat{W}_{j,-i}(k)$ is the $\hat{W}_j(k)$ computed with the $(n-1)$ observations after deleting the i-th observation from the original sample. Then, the corresponding jackknife pseudo-values are

$$W_i(k) = n\widehat{V} - (n-1)\widehat{V}_i,\ i = 1, \cdots, n.$$

Since the jackknife pseudo values are asymptotically independent under mild conditions (Tukey [72], Shi [73]), the standard empirical likelihood method could be applied to these jackknife samples for constructing an EL-based confidence interval for the CV. Let $\mathbf{P} = (p_1, ..., p_n)$ be a probability vector. The jackknife empirical likelihood for the CV can be defined as follows:

$$L_J(k) = \sup_{\mathbf{P}}\{\prod_{i=1}^{n} p_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i W_i(k) = 0\}.$$

Similarly, using the Lagrange multiplier method, we can get the Jackknife empirical

log-likelihood ratio for the CV:

$$l_J(k) = 2\sum_{i=1}^{n} log\{1 + \lambda_J W_i(k)\},$$

where λ_J is the solution of the equation:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{W_i(k)}{1 + \lambda_J W_i(k)} = 0.$$

Using similar techniques as in Jing *et al.* [40] and Peng [74], it can be proved that the Wilks' theorem for $l_J(k)$ still holds. i.e.,

$$l_J(k) \xrightarrow{d} \chi_1^2.$$

Therefore, a $100(1-\alpha)\%$ level JEL-based confidence interval for k can be constructed as follows:

$$\{k : l_J(k) \leq \chi_1^2(1-\alpha)\}.$$

5.4 Simulation Studies

5.4.1 Vangel's method

As we have discussed in the previous sections, due to the complicated distribution function for the sample CV, constructing an exact confidence interval for the CV become quite difficult. Therefore, approximation methods were proposed in the literature to avoid cumbersome calculations and intensive computations. The core part of approximation methods is selecting an appropriate pivot quantity. David [75] and Vangel[76] proposed their modified approximation methods based on different selections for pivot quantity. We provide the detailed procedure of Vangel's [76] work below to illustrate his method.

Vangel [76] defined a class of random variables as shown below:

$$Q = \frac{K^2(1+k^2)}{(1+\theta K^2)k^2},$$

where $\theta = \theta(k, \alpha)$ is a known function of k and α, $K = \frac{S}{\bar{X}}$ is the sample CV.

Suppose that Y_v is a random variable following the chi-square distribution with degrees of freedom $v = n - 1$, and $W_v = \frac{Y_v}{v}$. Let $t = \frac{\chi^2_{v,\alpha}}{v}$ denote the corresponding quantile of W_v where $\chi^2_{v,\alpha}$ is the 100α percentile of the distribution of Y_v, and $\alpha \in (0, 1)$. The approximation methods aim to select a proper θ such that $Pr(Q < t) \approx Pr(W_v \leq t)$. Vangel [76] proved that the distribution of W_v is known and is free of K, and he proposed a modified θ:

$$\theta = \frac{v}{v+1}[\frac{2}{\chi^2_{v,\alpha}} + 1].$$

Then, by using the approximate pivot, with a selected θ, a $100(1-\alpha)\%$ approximate confidence interval for k is constructed as follows:

$$\left(\frac{K}{\sqrt{t_1(\theta_1 K^2 + 1) - K^2}},\ \frac{K}{\sqrt{t_2(\theta_2 K^2 + 1) - K^2}} \right),$$

where $t_1 = \frac{\chi^2_{v,1-\frac{\alpha}{2}}}{v}$, $t_2 = \frac{\chi^2_{v,\frac{\alpha}{2}}}{v}$, and $\theta_i = \frac{2}{(v+1)t_i} + \frac{v}{k+1},\ i = 1, 2$. So, Vangel's interval is defined as:

$$\left(K[(\frac{u_1+2}{v+1} - 1)K^2 + \frac{u_1}{v}]^{-\frac{1}{2}},\ K[(\frac{u_2}{v+1} - 1)K^2 + \frac{u_2+2}{v}]^{-\frac{1}{2}} \right),$$

where $u_i \equiv vt_i$, for $i = 1, 2$.

However, there still exist some shortcomings with McKay and Vangel's methods. It was suggested by Mckay himself that his method could only be applied with small values for k, particularly for $k < 0.33$ (McKay, [77]). The pivotal quantity may be a complicated value when we select a large value for k which directly limit the application of these methods. So these problems strengthen our motivation to develop new non-parametric methods for construction of confidence interval for the CV.

5.4.2 Generalized Confidence Interval for a CV

In this section, we provide the detailed procedures to construct GPQ-based confidence intervals for a single CV when the underlying distribution F follows a normal distribution, a log-normal distribution and an Inverse Gaussian distribution. Inverse Gaussian distribution is an important family of skewed distributions which are often used to model income distributions in economics study.

Let χ be the sample space of possible values of X and x be an observed value of X. Let $R = r(X; x, \nu)$ be a function of (X, x, ν). R is a generalized pivotal quantity (GPQ) if it satisfies the following two conditions:

(a) It's distribution free of all the unknown parameters;

(b) The observed pivotal, $r_{obs} = r(x; x, \nu)$, does not depend on the nuisance parameter δ;

Then, without loss of generality, if $r_{obs} = \theta$, then a two-sided $(1-\alpha)$ level confidence interval for θ is provided by

$$(R_{\alpha/2}, R_{1-\alpha/2}).$$

where $R_{\alpha/2}$ and $R_{1-\alpha/2}$ are the $(\alpha/2)$-th and $(1-\alpha/2)$-th quantiles of the distribution of R. Weerahandi [52] and Hanning *et al.* [57] presented more details about generalized pivotal quantity and construction of GPQ-based confidence intervals.

Let $\{X_1, \cdots, X_n\}$ be a random sample following Inverse Gaussian distribution $IG(\mu, \lambda)$ with mean μ and scale parameter λ, the CV $k = \frac{\sqrt{\frac{\mu^3}{\lambda}}}{\mu} = \sqrt{\frac{\mu}{\lambda}}$. Then, the maximum likelihood estimators (MLEs) of μ and λ are $\hat{\mu} = \bar{X}, \hat{\lambda} = [\frac{1}{n}\sum_{i=1}^{n}(X_i^{-1} - \bar{X}^{-1})]^{-1}$ respectively, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

For notational convenience, we define $V = \hat{\lambda}^{-1}$. Then $\bar{X}$ and V are mutually independent random variables with distributions $\bar{X} \sim IG(\mu, n\lambda)$, $n\lambda V \sim \chi^2_{n-1}$. Let $\{x_1, ..., x_n\}$ be a given sample from $IG(\mu, \lambda)$. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $v = \frac{1}{n}\sum_{i=1}^{n}(x_i^{-1} - \bar{x}^{-1})$ are the observed values for $\bar{X}$ and V. The generalized pivot quantities for λ and μ (see Ye *et al.* [78]) are

given by

$$R_\lambda = \frac{n\lambda V}{nv} \sim \frac{\chi^2_{n-1}}{nv}, \tag{5.5}$$

$$R_\mu = \frac{\bar{x}}{|1+\frac{\sqrt{n\lambda(\bar{x}-\mu)}}{\mu\sqrt{\bar{x}}}\sqrt{\frac{\bar{x}}{nR_\lambda}}|} \overset{d}{\sim} \frac{\bar{x}}{|1+Z\sqrt{\frac{\bar{x}}{nR_\lambda}}|}, \tag{5.6}$$

respectively, where $Z \sim N(0,1)$, and $\overset{d}{\sim}$ denotes "approximately distributed" .

Therefore, the GPQ for the CV k is

$$R_k = \sqrt{\frac{R_\mu}{R_\lambda}}, \tag{5.7}$$

We propose the following Monte-Carlo algorithm for construction of a GPQ-based confidence interval for the CV:

STEP 1: Compute $\bar{x}$ and v using the original sample $\{x_1, ..., x_n\}$.

STEP 2: Generate one value of χ^2_{n-1} from the chi-squared distribution with $n-1$ degrees of freedom and one value of Z from the standard normal distribution.

STEP 3: Calculate R_λ and R_μ by using (5.5) and (5.6).

STEP 4: Calculate R_k by using (5.7).

STEP 5: Repeat STEPs 2-4 H times ($H \geq 10000$ is recommended) to obtain H copies of R_k: $\{R_{k,1}, \cdots, R_{k,H}\}$.

Then, a $100(1-\alpha)\%$ GPQ-based confidence interval for the CV can be constructed as

$$(R_{k,\alpha/2},\ R_{k,1-\alpha/2}),$$

where $R_{k,\alpha/2}$ and $R_{k,1-\alpha/2}$ are the $100(\alpha/2)$-th and $100(1-\alpha/2)$-th percentiles of $\{R_{k,1}, \cdots, R_{k,H}\}$, respectively.

Similarly, when the underlying distribution F follows a normal distribution with mean μ and standard deviation σ, the CV $k = \frac{\sigma}{\mu}$. From Tian [71], we can get a GPQ-based interval for k. Let $\{X_1, \cdots, X_n\}$ be a random sample from $N(\mu, \sigma^2)$. Then a sufficient estimator

for (μ, σ) is $(\hat{\mu}, \hat{\sigma}) = (\bar{X}, S)$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})$. The parameter of interest k can be estimated by $K = \frac{S}{\bar{X}}$. Given n observations $\{x_1, ..., x_n\}$ for $X \sim N(\mu, \sigma^2)$, the observed values of $\bar{X}$ and S^2 are $\bar{x}$ and s^2 respectively. Since

$$U \equiv \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}, \qquad Z \equiv (\frac{\sigma^2}{n})^{-1/2}(\bar{X} - \mu) \sim N(0, 1),$$

the GPQ's of σ^2 and μ are

$$R_{\sigma^2} = \frac{(n-1)s^2}{U}, \tag{5.8}$$

$$R_{\mu} = \bar{x} - (\frac{R_{\sigma^2}}{n})^{1/2} Z, \tag{5.9}$$

respectively.

Therefore, the GPQ for the CV k is

$$R_k = \frac{\sqrt{R_{\sigma^2}}}{R_{\mu}}. \tag{5.10}$$

Using the similar Monte-Carlo algorithm above (see also Tian [71]), we can get a $100(1-\alpha)\%$ GPQ-based confidence interval for the CV when the underlying distribution F follows a normal distribution.

5.4.3 Simulation Results

In this section, we conduct simulation studies to compare the newly proposed confidence intervals with existing confidence intervals in terms of coverage probability (CP) and average interval length (AL). The intervals being examined are Vangel's approximation-based interval, the bootstrap percentile (BP) interval, the proposed PEL interval, bootstrap-based PBEL interval, the Jackknife empirical likelihood-based JEL interval, the influence function-based IFEL interval. Another "exact" parametric confidence interval, called *Generalized Pivotal Quantity* (GPQ) based interval, is also included for comparison.

We carry out the simulation studies with sample size n = 50, 300, 500, 800, 1000,

respectively, to examine the finite sample performances of the various intervals for the CV from smaller sample size to larger sample size. $R = 1000$ iterations is chosen for calculating coverage probabilities and average interval lengths of various intervals. $B = 1000$ bootstrap replications is chosen for calculating BPEL and BP intervals. $H = 10000$ replications is chosen for calculating GPQ-based intervals. Since a larger value of the coefficient of variation results in bigger relative variation which is of less interest in practice, and Miller [79] indicated that for normally distributed data encountered in real-life settings, it is not expected that the CV value will exceed 0.33, we select the true CV values k =0.2, 0.5 to investigate the performance of all methods mentioned above.

Four underlying distributions, a normal distribution, a log-normal distribution, an Inverse Gaussian distribution and a Chi-square distribution, are selected to generate random samples. For the setting with the normal distribution as the underlying distribution, we consider all the methods mentioned above. When the underlying distribution is a log-normal distribution or an Inverse Gaussian distribution, Vangel's method cannot be applied to, but we can apply the GPQ-based method (because $log(X) \sim N(\mu, \sigma^2)$ for a log-normal distribution) and the non-parametric methods. In the other setting, the Chi-square distribution is selected as the underlying distribution to examine the BP interval and the proposed non-parametric intervals. Since the normality assumption is invalid here, Vangel and GPQ-based intervals cannot be applied to this setting.

At 90% confidence level, we calculate the coverage probabilities that the confidence intervals cover the true value of k. The closer the coverage probability is to the nominal level, the better performance of the confidence interval. With similar coverage probabilities, the shorter average length is, the better performance of the confidence interval. The simulation results are displayed in Tables 5.1-5.4. From Tables 5.1-5.3, we observe that the GPQ-based interval has the best performance among all intervals regardless of the sample size when the underlying distribution is a normal distribution or a log-normal distribution or an Inverse Gaussian distribution. Its coverage probabilities are very close to the nominal level. When the underlying distribution is a normal distribution, the parametric Vangel intervals have

good coverage accuracy when $n \geq 500$; BP interval and all the EL-based (PEL, BPEL, JEL and IFEL) intervals have acceptable coverage probabilities as sample size n increases to 300 except the PEL and BPEL intervals that have over-coverage problems when $k = 0.2$. When the underlying distribution is a log-normal distribution and k is small (i.e., $k = 0.2$), JEL and IFEL intervals have acceptable coverage probabilities as sample size n increases to 300; when $k = 0.5$, BP interval and all the EL-based intervals have under-coverage problems. When the underlying distribution is an Inverse Gaussian distribution, BP interval and all the EL-based (PEL, BPEL, JEL and IFEL) intervals have acceptable coverage probabilities as sample size n increases to 500 except the BPEL intervals that have over-coverage problems when $k = 0.2$. From Table 5.4, we can see that BP interval, BPEL and JEL intervals undercover the true value of k when $k = 0.5$ with small to moderate sample sizes ($n = 50, 300, 800$), but the coverage probabilities are closer to the nominal level as sample size n increases to 1000. BP interval and all the EL-based (PEL, BPEL, JEL and IFEL) intervals have coverage probabilities close to the nominal level as sample size $n \geq 300$ when $k = 0.2$ except PEL has over-coverage problems.

Above all, under the normality/log-normality/Inverse Gaussian assumptions, the GPQ-based intervals have the best performance with good coverage probability and stability. Therefore, we recommend the GPQ-based intervals for the CV when the underlying distribution is normal/log-normal/Inverse Gaussian. When the underlying distribution is a non normal/log-normal/Inverse Gaussian or an unknown distribution, BP interval and the EL-based intervals for the CV have acceptable coverage accuracy when sample size is big enough. Since the bootstrap-based BP interval is computationally more intensive than the EL-based intervals when sample size is big, EL-based intervals for the CV are preferred when the underlying distribution is unknown and sample size is acceptably large.

Table (5.1) Coverage probabilities (CP) and average lengths (AL) of various 90% level confidence intervals for the CV. Underlying distribution: Normal distribution $N(1, k)$.

k	n	Method	Vangel	GPQ	BP	PEL	BPEL	JEL	IFEL
0.2	50	CP	0.880	0.902	0.872	0.870	0.962	0.874	0.880
		AL	0.065	0.071	0.065	0.071	0.095	0.068	0.072
	300	CP	0.894	0.902	0.890	0.940	1.000	0.900	0.900
		AL	0.027	0.028	0.027	0.030	0.090	0.028	0.034
	500	CP	0.893	0.897	0.885	0.940	1.000	0.895	0.910
		AL	0.021	0.022	0.021	0.025	0.071	0.022	0.025
	800	CP	0.897	0.903	0.885	0.890	1.000	0.899	0.910
		AL	0.017	0.017	0.017	0.019	0.062	0.017	0.019
	1000	CP	0.899	0.902	0.891	0.870	1.000	0.902	0.890
		AL	0.015	0.015	0.015	0.017	0.060	0.015	0.016
0.5	50	CP	0.887	0.903	0.873	0.940	0.801	0.828	0.880
		AL	0.206	0.215	0.195	0.242	0.157	0.191	0.194
	300	CP	0.912	0.914	0.894	0.930	0.896	0.891	0.900
		AL	0.083	0.083	0.081	0.115	0.080	0.084	0.108
	500	CP	0.899	0.898	0.891	0.920	0.891	0.874	0.920
		AL	0.065	0.064	0.063	0.076	0.063	0.065	0.080
	800	CP	0.895	0.891	0.891	0.890	0.876	0.877	0.900
		AL	0.051	0.051	0.050	0.053	0.044	0.048	0.069
	1000	CP	0.901	0.897	0.893	0.920	0.892	0.871	0.900
		AL	0.046	0.045	0.045	0.047	0.045	0.042	0.060

Table (5.2) Coverage probabilities (CP) and average lengths (AL) of various 90% level confidence intervals for the CV. Underlying distribution: Log-N$(1, [log(k^2+1)]^2)$.

k	N	Method	GPQ	BP	PEL	BPEL	JEL	IFEL
0.2	50	CP	0.905	0.855	0.830	0.876	0.861	0.850
		AL	0.070	0.064	0.069	0.067	0.070	0.068
	300	CP	0.902	0.879	0.950	0.987	0.891	0.920
		AL	0.028	0.028	0.033	0.047	0.030	0.031
	500	CP	0.889	0.891	0.950	0.999	0.901	0.920
		AL	0.021	0.022	0.027	0.046	0.024	0.024
	800	CP	0.890	0.874	0.900	1.000	0.899	0.880
		AL	0.017	0.018	0.022	0.047	0.020	0.019
	1000	CP	0.904	0.886	0.910	1.000	0.924	0.870
		AL	0.015	0.016	0.019	0.046	0.018	0.017
0.5	50	CP	0.905	0.779	0.820	0.780	0.789	0.780
		AL	0.193	0.177	0.164	0.174	0.210	0.153
	300	CP	0.902	0.837	0.830	0.848	0.846	0.890
		AL	0.076	0.091	0.093	0.094	0.100	0.096
	500	CP	0.889	0.854	0.890	0.858	0.850	0.880
		AL	0.058	0.073	0.071	0.076	0.079	0.076
	800	CP	0.890	0.845	0.890	0.856	0.844	0.890
		AL	0.046	0.059	0.060	0.060	0.063	0.065
	1000	CP	0.904	0.859	0.860	0.860	0.852	0.850
		AL	0.041	0.053	0.054	0.054	0.056	0.060

Table (5.3) Coverage probabilities (CP) and average lengths (AL) of various 90% level confidence intervals for the CV. Underlying distribution: Inverse Gaussian distribution $IG(1, \frac{1}{k^2})$.

k	n	Method	GPQ	BP	PEL	BPEL	JEL	IFEL
0.2	50	CP	0.909	0.870	0.850	0.870	0.870	0.890
		AL	0.069	0.063	0.064	0.058	0.069	0.066
	300	CP	0.897	0.890	0.910	0.990	0.890	0.930
		AL	0.027	0.028	0.029	0.073	0.029	0.030
	500	CP	0.887	0.890	0.900	0.980	0.910	0.920
		AL	0.021	0.022	0.023	0.060	0.023	0.024
	800	CP	0.893	0.880	0.890	1.000	0.920	0.910
		AL	0.017	0.017	0.020	0.063	0.019	0.020
	1000	CP	0.901	0.890	0.860	1.000	0.940	0.870
		AL	0.015	0.015	0.016	0.066	0.018	0.018
0.5	50	CP	0.907	0.800	0.790	0.720	0.820	0.820
		AL	0.181	0.172	0.159	0.136	0.203	0.161
	300	CP	0.890	0.870	0.920	0.840	0.860	0.910
		AL	0.072	0.085	0.088	0.077	0.091	0.095
	500	CP	0.890	0.870	0.890	0.860	0.870	0.920
		AL	0.055	0.066	0.070	0.064	0.070	0.075
	800	CP	0.888	0.860	0.870	0.880	0.870	0.900
		AL	0.044	0.053	0.057	0.056	0.056	0.065
	1000	CP	0.899	0.900	0.880	0.930	0.900	0.910
		AL	0.039	0.048	0.050	0.055	0.051	0.061

Table (5.4) Coverage probabilities (CP) and average lengths (AL) of various 90% level confidence intervals for the CV. Underlying distribution: Chi-square distribution $\chi^2_1(\frac{2}{k^2})$.

k	n	Method	BP	PEL	BPEL	JEL	IFEL
0.2	50	CP	0.870	0.920	0.815	0.880	0.920
		AL	0.062	0.073	0.052	0.066	0.067
	300	CP	0.894	0.950	0.911	0.909	0.890
		AL	0.027	0.032	0.027	0.028	0.028
	500	CP	0.897	0.940	0.903	0.904	0.900
		AL	0.021	0.024	0.021	0.021	0.022
	800	CP	0.895	0.940	0.898	0.906	0.920
		AL	0.017	0.020	0.017	0.017	0.017
	1000	CP	0.904	0.940	0.908	0.909	0.890
		AL	0.015	0.017	0.015	0.015	0.015
0.5	50	CP	0.839	0.880	0.809	0.744	0.860
		AL	0.161	0.167	0.140	0.150	0.159
	300	CP	0.873	0.910	0.862	0.851	0.880
		AL	0.072	0.077	0.070	0.070	0.079
	500	CP	0.889	0.900	0.888	0.878	0.900
		AL	0.056	0.062	0.056	0.056	0.063
	800	CP	0.871	0.920	0.873	0.875	0.930
		AL	0.045	0.048	0.044	0.045	0.050
	1000	CP	0.897	0.900	0.889	0.904	0.940
		AL	0.040	0.043	0.040	0.041	0.044

5.5 Real Examples

In this section, two real examples are used to illustrate the methods proposed in the previous sections.

The Beef Council Check-off dataset is obtained from The Data and Story Library (DASL) at Carnegie Mellon University. The US Congress created the Beef Promotion and Research Act, the "Beef Checkoff Program", with passage of the 1985 Farm Bill. By law, all producers selling cattle or calves must pay $1 per head to support beef/veal promotion, research and information. We select the average value of products sold (thousands) as the underlying variable of interest in the dataset. There are 56 observations for the selected variable. The estimated (sample) coefficient of variation of the average value of products

sold is $K = 0.4733$. To construct a confidence interval for the CV, the Shapiro-Wilk test (Shapiro and Wilk, [61]) is conducted to check for the normality of the distribution of the variable. The resulting Shapiro-Wilk statistic is associated with a p-value of 0.5037 which cannot reject the null hypothesis that the dataset is normally distributed. With our simulation results in mind, we can now apply the GPQ-based method. The 90% level GPQ-based interval for the CV is (0.3973, 0.5872), which indicates that the average values of products sold have small relative-variation. Therefore, we can conclude that the Beef products sold in US are of good quality.

The second dataset comes from The Panel Study of Income Dynamics (PSID) which is a longitudinal survey of the families in USA. The PSID has been conducted by the University of Michigan since 1968 and data about US families are collected annually. We use the family income data from year 2000 from the PSID Family Income Plus Files. The dataset consists of 7,406 families. The sample coefficient of variation for the family income is $K = 0.6535$. Obviously, the family income distribution in the US is a skewed (i.e. non-normal) distribution. Therefore, we can apply the newly proposed non-parametric methods to calculate confidence intervals for the CV. The 90% level BP interval and EL-based confidence intervals for the CV are displayed in Table 5.5. From Table 5.5, we can see that BP method and all the EL-based methods produce similar confidence intervals in this example. In particular, both BPEL and IFEL methods give the same interval for the CV which is (0.585, 0.722). This interval shows that the relative variation of family income around the mean family income in USA is at a moderate level. This finding will provide meaningful information on income inequality for US government.

Table (5.5) 90% level confidence intervals for the CV of the family income in USA

K	Method	BP	PEL	BPEL	JEL	IFEL
0.6535	Upper bound	0.588	0.655	0.585	0.599	0.585
	Lower bound	0.720	0.722	0.722	0.749	0.722

PART 6

CONCLUSIONS AND FUTURE STUDIES

In this dissertation, we generalize the approach of Faraggi [9] and propose HWS, HAC and HBCA intervals for the covariate-adjusted Youden index under the heterosedastic regression models with/without the normality assumption for the error distributions. Meanwhile, we developed influence function based empirical likelihood (IFEL) method to construct confidence intervals/regions for the parameters in the AUC regression along with the correlation coefficient (CC) and the coefficient of variation (CV), respectively. Through the intensive simulation studies, we can draw the conclusions as follows.

First, we proposed various confidence intervals for the covariate-adjusted Youden index along with the optimal cut-off point under linear regression model and heterosedastic regression model. Our simulation results have shown that with normal error assumption, the GPQ-based intervals outperform the bootstrap-based BCa, BTI and BTII intervals under the same parametric linear models setting, particularly for small to moderate sized samples which are more applicable and practical in second or third phase medical diagnostic studies. Under heterosedastic regression, we apply the MOVER method to construct the hybrid CI for the covariate-adjusted YI with/without normal error assumption. When the errors are normally distributed, HBCA-N method outperforms HAC-N, HWS-N and ACNA method in terms of coverage probabilities. HAC-N intervals performs better than HWS-N intervals. When the errors are not normally distributed, HBCA-E and ACNA intervals for the covariate-adjusted Youden Index are farther from the 95% nominal level at some values of covariates than HAC-E and HWS-E intervals. And HAC-E intervals have better performance and are much more stable than HWS-E intervals in terms of coverage probabilities. Without the normal error assumptions, we also examine the performance of HBCA-E and ACNA intervals for the optimal cut-off point at given $Z = z$. ACNA intervals are much

more stable than HBCA-E intervals, although ACNA intervals are slightly higher than the 95% nominal level at the majority values of covariates. Above all, under the heteroscedastic regression models, we recommend the HAC-E intervals for the covariate-adjusted YI and ACNA intervals for the optimal cut-off point for the test results in practice.

In ROC analysis, the area under the ROC curve (AUC) is a popular summary measurement of the discriminatory accuracy of a diagnostic test. AUC Regression is commonly used to evaluate the effects of the covariates on the diagnostic accuracy. Since the asymptotic distribution of the influence function-based empirical log-likelihood ratio statistic is a standard chi-square distribution. Hence, confidence regions based on the influence function for the regression parameters can be easily obtained without any variance estimates. Hereby, in Part 3, we develop new EL-based statistical methods for the AUC regression including Bootstrap influence function-based method (BIFEL) and Jackknife empirical likelihood-based method (JEL). We compare our proposed methods with the existing method by constructing confidence regions for inferences of the AUC regression. Simulation studies indicate BIFEL confidence regions outperform JEL and NA confidence regions in terms of coverage probabilities. Finally, a real study of the distortion product otoacoustic emissions (DPOAE) test to diagnose the hearing impairment demonstrates the application of our proposed method.

In Part 4, we proposed an influence function-based empirical likelihood method to construct a confidence interval for the CC. The simulation results indicate that the GPQ-based interval performs very well when the underlying distribution is normal while the IFEL interval has better overall performances with finite samples when the underlying distribution is non-normal or unknown. The proposed EL-based intervals are easy to use and can be directly calculated by implementing the algorithm for computing the standard empirical likelihood interval (Hall and La Scala [14]). We recommend the use of the z-transformation based NAI interval, the GPQ interval, and the IFEL interval when the underlying distribution is a normal distribution, and the use of the IFEL interval when the underlying distribution is non-normal or unknown.

In Part 5, we proposed an empirical likelihood method based on influence function

to construct confidence intervals for the CV. Simulation studies showed that under the normality/log-normality assumptions, the GPQ-based interval has the best performance with good coverage probability and stability. When the underlying distribution is a non normal/log-normal or an unknown distribution, BP interval and the EL-based intervals for the CV have acceptable coverage accuracy when sample size is acceptably large ($n \geq 1000$). Thus, these non-parametric methods should also be considered for obtaining confidence intervals for the CV when the underlying distribution is unknown. In terms of computation time, EL-based methods are more efficient than BP-based method especially for large sample sizes. Hereby, we recommend using GPQ-based interval for the CV when the underlying distribution is normal/log-normal and EL-based interval when the underlying distribution is not normal or unknown and sample size is acceptably large.

With the highly developed medical technologies, for some diseases, eg. Brain tumor, Alzheimer's Disease, a transitional stage, which is between healthy and unhealthy stage, could be detected and defined. In order to gain the optimal timing window for medical interventions, it is necessary to recognize the intermediate stage if it exists. In the future studies, we may consider extending the original two groups setting into three ordinal groups setting and redefining the sensitivity, specificity and adding the transitional rate. Then, the three-dimensional ROC surface will be the measurement of correct classification into three diagnostic groups used for testing accuracy. It's necessary and valuable for researchers to think about how to apply the influence function-based empirical likelihood method to construct confidence intervals for adjusted YI, optimal cutoff points and make inferences for the volume under ROC surface (VUS) in the presence of covariates. Do the proposed non-parametric empirical estimation methods still perform well for the three-groups setting ? Also, in part two, without the normal error assumptions, for the non-parametric empirical estimation method, we only consider the t distribution as an example. How about application for the skewed data and the longitudinal data? All above questions and topics are critical for the future empirical methods studies in biomedical research and other scientific disciplines, specifically for interval estimation of parameters of interest.

REFERENCES

[1] W. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, pp. 32–35, 1950.

[2] Pardo-Fernandez, M. J.C., Rodriguez-Alvarez, and I. Van Keilegom, "A review on roc curves in the presence of covariates," *REVSTAT Statistical Journal*, vol. 1, pp. 21–41, 2014.

[3] M. S. Pepe, "Three approaches to regression analysis of receiver operating characteristic curves for continuous test results," *Biometrics*, vol. 54, pp. 124–135, 1998.

[4] A. N. A. Tosteson and C. B. Begg, "A general regression methodology for roc curve estimation," *Medical Decision Making*, vol. 8, pp. 204–215, 1988.

[5] P. Smith and T. Thompson, "Correcting for confounding in analyzing receiver operating characteristic curves," *Biometrical Journal*, vol. 38, pp. 857–863, 1996.

[6] X. H. Zhou, N. A. Obuchowski, and D. K. McClish, *Statistical methods in diagnostic medicine*, 2009.

[7] Y. Zheng and P. J. Heagerty, "Semiparametric estimation of time-dependent roc curves for longitudinal marker data," *Biostatistics*, vol. 5, pp. 615–632, 2004.

[8] A. Rodríguez and J. C. Martínez, "Bayesian semiparametric estimation of covariate-dependent roc curves," *Biostatistics*, vol. 15, pp. 353–369, 2014.

[9] D. Faraggi, "Adjusting receiver operating characteristic curves and related indices for covariates," *Journal of the Royal Statistical Society: Series D (the Statistician)*, vol. 52, pp. 179–192, 2003.

[10] F. Yao, R. V. Craiu, and B. Reiser, "Nonparametric covariate adjustment for receiver operating characteristic curves," *Can. J. Stat.*, vol. 38, pp. 27–46, 2010.

[11] H. Zhou and G. Qin, "Nonparametric covariate adjustment for the youden index," *Applied Statistics in Biomedicine and Clinical Trials Design*, pp. 109–132, 2015.

[12] A. B. Owen, "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, vol. 75, pp. 237–249, 1988.

[13] ——, "Empirical likelihood ratio confidence regions," *Biometrika*, vol. 18, pp. 90–120, 1990.

[14] P. Hall and B. La Scala, "Methodology and algorithms of empirical likelihood," *International Stiatiscal Review*, vol. S8, 2, pp. 109–127, 1990.

[15] O. A., *Empirical likelihood*, 2001.

[16] G. Claeskens, B. Jing, L. Peng, and W. Zhou, "Empirical likelihood confidence regions for comparison distributions and roc curves," *Canad. J. Statist.*, vol. 31, pp. 173–190, 2003.

[17] G. Qin and X. H. Zhou, "Empirical likelihood inference for the area under the roc curve," *Biometrics*, vol. 62, pp. 613–622, 2006.

[18] M. S. Pepe, "The statistical evaluation of medical tests for classification and prediction," 2003.

[19] K. W. Tsui and S. Weerahandi, "Generalized p-values insignificance testing of hypotheses in the presence of nuisance parameters," *Journal of American Statistical Association*, vol. 84, pp. 602–607, 1989.

[20] J. Gamage, T. Mathew, and S. Weerahandi, "Generalized p-values and generalized confidence regions for the multivariate behrens-fisher problem and manova," *Journal of Multivariate Analysis*, vol. 88, no. 177-189, 2004.

[21] J. C. Lee and S. H. Lin, "Generalized confidence intervals for the ratio of means of two normal populations," *Journal of Statistical Planning and Inference*, vol. 123, pp. 49–60, 2004.

[22] L. Tian and G. E. Wilding, "Confidence interval estimation of a common correlation coefficient," *Computational Statistics and Data Analysis*, vol. 52, pp. 4872–4877, 2008.

[23] L. Tian, "Confidence interval estimation of a common correlation coefficient," *Computational Statistics & Data Analysis*, vol. 52, pp. 4872–4877, 2008.

[24] C. Y. Lai, L. Tian, and E. F. Schisterman, "Exact confidence interval estimation for the youden index and its corresponding optimal cut-off point," *Computational Statistics & Data Analysis*, vol. 56, pp. 1103–1114, 2012.

[25] S. Weerahandi, "Generalized confidence intervals," *Journal of American Statistical Association*, vol. 88, pp. 899–905, 1993.

[26] ——, "Exact methods in manova and mixed models," *Generalized inference in repeated measures*, 2004.

[27] ——, *Exact statistical methods for data analysis*, 2013.

[28] V. Inácio de Carvalho, M. de Carvalho, and A. J. Branscum, "Nonparametric bayesian covariate-adjusted estimation of the youden index," *Biometrics*, vol. 73, pp. 1279–1288, 2017.

[29] E. F. Schisterman and N. Perkins, "Confidence intervals for the youden index and corresponding optimal cut-off point," *Communications in Statistics, Simulation and Computation*, vol. 36, pp. 549–563, 2007.

[30] N. W. Hengartner, M. H. Wegkamp, and E. Matzner-Løber, "Bandwidth selection for local linear regression smoothers," *Journal of the Royal Statistical Society: Series B*, vol. 64, pp. 791–804, 2002.

[31] J. Fan and I. Gijbels, *Local Polynomial Modelling and Its Applications*, 1996.

[32] G. Y. Zou, "Confidence interval estimation for lognormal data with application to health economic," *Computational Statistics & Data Analysis*, vol. 53, pp. 3755–3764, 2009.

[33] M. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, 2003.

[34] X. Zhou, N. Obuchowski, and D. McClish, "Statistical methods in diagnostic medicine," 2002.

[35] M. Thompson and W. Zucchini, "On the statistical analysis of roc curves," *Statistics in Medicine*, vol. 8, pp. 1277–1290, 1989.

[36] N. Obuchowski, "Multireader, multimodality receiver operating characteristic curve studies: Hypothesis testing and sample size estimation using analysis of variance approach with dependent observations," *Academic Radiology*, vol. 2, no. Suppl.1, pp. S22–S29, 1995.

[37] D. Dorfman, K. Berbaum, and C. Metz, "Receiver operating characteristic analysis: generalization to the population of readers and patients with the jackknife method." 1992.

[38] L. Dodd and M. Pepe, "Partial auc estimation and regression," *Biometrics*, vol. 59, pp. 614–623, 2003.

[39] M. Pepe and T. Cai, "The analysis of placement values for evaluating discriminatory measures," *Biometrics*, vol. 60, pp. 528–535, 2004.

[40] B. Y. Jing, J. Yuan, and W. Zhou, "Jackknife empirical likelihood," *Journal of the American Statistical Association*, vol. 195, pp. 1224–1232, 2009.

[41] X. Zhou, G. Qin, H. Lin, and G. Li, "Inferences in censored cost regression models with empirical likelihood," *Statistica Sinica*, vol. 16, pp. 1213–1232, 2006.

[42] L. Stover, M. Gorga, S. Neely, and D. Montoya, "Toward optimizing the clinical utility of distortion product otoacoustic emission measurements," *Journal of the Acoustical Society of America*, vol. 100, pp. 956–967, 1996.

[43] F. Galton, "Regression towards mediocrity in hereditary stature," *Journal of the Anthropological Institute*, vol. 15, pp. 246–263, 1885.

[44] K. Pearson, "Royal society proceedings," vol. 58, p. 241, 1895.

[45] ——, "Notes on the history of correlation," *Biometrika*, vol. 13, pp. 25–45, 1920.

[46] J. L. Rogers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *The American Statistician*, vol. 42, pp. 59–66, 1988.

[47] R. Fisher, "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," *Biometrika*, vol. 10, pp. 507–521, 1915.

[48] H. Hotelling, "New light on the correlation coefficient and its transform," *Journal of the Royal Statistical Society, Serie B*, vol. 15, pp. 193–232, 1953.

[49] R. Fisher, "On the 'probable error' of a coefficient of correlation deduced from a small sample," *Metron*, vol. 1, pp. 3–32, 1921.

[50] G. J. Weerakkody and S. Givaruangsawat, "Estimating the correlation coefficient in the presence of correlated observations from a bivariate normal population," *Communication in Statistics-Theory and Methods*, vol. 24, pp. 1705–1719, 1995.

[51] Y. Sun and A. C. M. Wong, "Interval estimation for the normal correlation coefficient," *Statistic & Probability Letters*, vol. 77, pp. 1652–1661, 2007.

[52] S. Weerahandi, "Generalized confidence intervals," *Journal of the American Statistical Association*, vol. 88, pp. 899–905, 1993.

[53] L. Nie, Y. Chen, and H. Chu, "Asymptotic variance of maximum likelihood estimator for the correlation coefficient from a bvn distribution with one variable subject to censoring," *Journal of Statistical Planning and Inference*, vol. 141, pp. 392–401, 2011.

[54] J. E. Freud, "A bivariate extension of the exponential distribution," *Journal of the American Statistical Association*, vol. 56, pp. 971–977, 1961.

[55] F. Downton, "Bivariate exponential distributions in reliability theory," *Journal of the Royal Statistical Society, Serie B*, vol. 32, pp. 408–417, 1970.

[56] S. Al-Saadi and D. Young, "Estimators for the correlation coefficient in a bivariate exponential distribution," *Journal of Statistical Computation and Simulation*, vol. 28, pp. 13–20, 1980.

[57] J. Hanning, H. Iyer, and P. Patterson, "Fiducial generalized confidence intervals," *Journal of the American Statistical Association*, vol. 101, pp. 254–269, 2006.

[58] A. W. Marshall and I. Olkin, "A multivariate exponential distribution," *Journal of the American Statistical Association*, vol. 62, pp. 30–44, 1967.

[59] L. Willerman, R. J. N. Schultz, R., and E. Bigler, "In vivo brain size and intelligence," *Intelligence*, vol. 15, pp. 223–228, 1991.

[60] D. Wechsler, "Manual for the wechsler adult intelligence scale-revised," *Psychological Corporation*, 1981.

[61] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, pp. 591–611, 1965.

[62] M. MacGregor and E. Price, "exploration of aspects of language proficiency and algebra learning," *Journal of Research in Mathematics Education*, vol. 30, pp. 449–467, 1999.

[63] G. Taye and P. Njuho, "Monitoring field variability using confidence interval for coefficient of variation," *Communications in Statistics-Theory and Methods*, vol. 37, pp. 831–846, 2008.

[64] W. A. Hendricks and K. W. Robey, "The sampling distribution of the coefficient of variation," *Ann. Math. Statist.*, vol. 7, pp. 129–132, 1936.

[65] E. L. Lehmann, *Testing statistical hypothesis*, 1986.

[66] A. J. Hayter, "Confidence bounds on the coefficient of variation of a normal distribution with applications to win-probabilities," *Journal of Statistical Computation and Simulation*, vol. 18, pp. 3778–3791, 2015.

[67] A. Hayter and J. Kim, "Small sample tests for the equality of two normal cumulative probabilities, coefficients of variations and sharpe ratios," *Journal of Statistical Theory and Practice*, vol. 9, pp. 23–36, 2015.

[68] R. Mahmoudvand and H. Hassani, "Two new confidence intervals for the coefficient of variation in a normal distribution," *Journal of Applied Statistics*, vol. 36, pp. 429–442, 2009.

[69] W. Panichkitkosolkul, "Confidence intervals for the coefficient of variation in a normal distribution with a known population mean," *Journal of Probability and Statistics*, 2013.

[70] S. Verrill, "Confidence bounds for normal and lognormal distribution coefficients of variation," 2003.

[71] L. Tian, "Inferences on the common coefficient of variation," *Statistics in Medicine*, vol. 24, pp. 2213–2220, 2005.

[72] J. W. Tukey, "Bias and confidence in not-quite large samples," *Ann. Statist.*, vol. 29, p. 614, 1958.

[73] X. Shi, "The approximate independence of jackknife pseudo-values and the bootstrap methods," *Journal of Wuhan Institute Hydra-Electric Engineering*, vol. 2, pp. 83–90, 1984.

[74] L. Peng, "Approximate jackknife empirical likelihood method for estimating equations," *Canadian Journal of Statistics*, vol. 40, pp. 110–123, 2012.

[75] F. N. David, "Note on the application of fisher's k-statistics," *Biometrika*, vol. 36, pp. 383–393, 1949.

[76] M. G. Vangel, "Confidence intervals for a normal coefficient of variation," *Am. Statist.*, vol. 50, pp. 21–26, 1996.

[77] A. T. McKay, "Distribution of the coefficient of variation and the extended t-distribution," *J. Roy. Statist. Soc. B*, vol. 95, pp. 695–698, 1932.

[78] R. Ye, T. Ma, and S. Wang, "Inferences on the common mean of several inverse gaussian populations," *Comput. Statist. Data. Anal.*, vol. 54, pp. 906–915, 2010.

[79] G. E. Miller, "Asymptotic test statistics for coefficients of variation," *Commun. Statist. Theor. Meth*, vol. 20, pp. 3351–3363, 1991.

Appendix A

PROOFS OF PART 3

We need Lemma 3.1 and Lemma 3.2 for the proof of Theorem 3.1.

We denote $x^{\otimes 2} = xx^T$, for $x \in R^p$, and denote $\|\cdot\|$ the Euclidean norm.

Lemma 3.1 *Under the conditions in Theorem 1, we have*

$$\frac{1}{\sqrt{n+m}} \sum_{k=1}^{n+m} W_k(\beta) \xrightarrow{\mathcal{L}} N(0, \frac{\rho}{1+\rho} V).$$

where

$$\begin{aligned} V &= V_1 + V_2, \\ V_1 &= \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} (\omega(\beta_0^T \mathbf{Z}_j^D)\mathbf{Z}_j^D)^{\otimes 2} Var^D_{\mathbf{Z^D}}[F^{\bar{D}}_{\mathbf{Z^{\bar{D}}}}(Y^D)], \\ V_2 &= \rho(\lim_n \frac{1}{n} \sum_{j=1}^{n} \omega(\beta_0^T \mathbf{Z}_j^D)\mathbf{Z}_j^D)^{\otimes 2} Var^{\bar{D}}_{\mathbf{Z^{\bar{D}}}}[\int I(Y^{\bar{D}} \le t) dF^D_{\mathbf{Z^D}}(t)]. \end{aligned}$$

Proof. From

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \widehat{\mathbf{H}}_j = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [1 - \widehat{U}_j^D - g(\beta_0^T \mathbf{Z}_j^D)] \omega(\beta_0^T \mathbf{Z}_j^D) \mathbf{Z}_j^D,$$

and

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \mathbf{H_j} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [1 - U_j^D - g(\beta_0^T \mathbf{Z}_j^D)] \omega(\beta_0^T \mathbf{Z}_j^D) \mathbf{Z}_j^D,$$

we obtain the following decomposition:

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \widehat{\mathbf{H}}_j = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \mathbf{H_j} + \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(U_j^D - \widehat{U}_j^D \right) \omega(\beta_0^T \mathbf{Z}_j^D) \mathbf{Z}_j^D \equiv I_1 + I_2.$$

From

$$\begin{aligned} Var(I_1) &= \frac{1}{n} Var(\sum_{j=1}^{n} [F^{\bar{D}}_{\mathbf{Z}_{\bar{\mathbf{D}}}}(Y_j^D) - g(\beta_0^T \mathbf{Z}_j^D)] \omega(\beta_0^T \mathbf{Z}_j^D) \mathbf{Z}_j^D) \\ &= \frac{1}{n} \sum_{j=1}^{n} (\omega(\beta_0^T \mathbf{Z}_j^D) \mathbf{Z}_j^D)^{\otimes 2} Var^D_{\mathbf{Z^D}} [F^{\bar{D}}_{\mathbf{Z}_{\bar{\mathbf{D}}}}(Y_j^D)] \\ &\longrightarrow V_1 \end{aligned}$$

(If $\mathbf{Z}_j^D$'s are i.i.d. random variables, then $V_1 = E((\omega(\beta_0^T \mathbf{Z}^D)\mathbf{Z}^D)^{\otimes 2} Var^D_{\mathbf{Z^D}}[F^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(Y^D)]))$, and Central Limit Theorem, it follows that $I_1 \xrightarrow{\mathcal{L}} N(0, V_1)$.

For the term I_2, using $\widehat{F}^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(t) = \frac{1}{m} \sum_{i=1}^{m} I(Y_j^{\bar{D}} \le t)$, we get that

$$\begin{aligned} I_2 &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(\widehat{F}^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(Y_j^D) - F^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(Y_j^D) \right) \omega(\beta_0^T \mathbf{Z}_j^D) \mathbf{Z}_j^D \\ &= [\lim_n \frac{1}{n} \sum_{j=1}^{n} \omega(\beta_0^T \mathbf{Z}_j^D) \mathbf{Z}_j^D] \int \sqrt{n} \left(\widehat{F}^{\bar{D}}_{\bar{D}}(t) - F^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(t) \right) dF^D_{\mathbf{Z^D}}(t) + o_p(1) \\ &= \sqrt{\frac{n}{m}} [\lim_n \frac{1}{n} \sum_{j=1}^{n} \omega(\beta_0^T \mathbf{Z}_j^D) \mathbf{Z}_j^D] \\ &\quad \times \sqrt{m} \left(\frac{1}{m} \sum_{i=1}^{m} \int I(Y_i^{\bar{D}} \le t) dF^D_{\mathbf{Z^D}}(t) - \int F^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(t) dF^D_{\mathbf{Z^D}}(t) \right) + o_p(1) \\ &\xrightarrow{\mathcal{L}} N(0, V_2). \end{aligned}$$

For given covariates $\mathbf{Z_j^D}$'s, test results Y_j^D's for the diseased group and $Y_i^{\bar{D}}$'s for the non-diseased group are independent, so I_1 and I_2 are asymptotically independent. Therefore,

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \widehat{\mathbf{H}}_j = I_1 + I_2 \xrightarrow{\mathcal{L}} N(0, V). \tag{A.1}$$

with $V = V_1 + V_2$.

Observing that

$$\frac{1}{\sqrt{n+m}}\sum_{j=1}^{n}\widehat{\mathbf{H}}_j = \frac{1}{\sqrt{n+m}}\sum_{j=1}^{n}\mathbf{H_j} + \frac{n}{m}[\lim_n \frac{1}{n}\sum_{j=1}^{n}\omega(\beta^T\mathbf{Z}_j^D)\mathbf{Z}_j^D] \quad (A.2)$$

$$\times\frac{1}{\sqrt{n+m}}\sum_{i=1}^{m}\left(\int I(Y_i^{\bar{D}}\le t)dF_{\mathbf{Z^D}}^D(t) - \int F_{\mathbf{Z^{\bar{D}}}}^{\bar{D}}(t)dF_{\mathbf{Z^D}}^D(t)\right) + o_p(1) \quad (A.3)$$

$$= \frac{1}{\sqrt{n+m}}\sum_{j=1}^{n}\mathbf{H_j} + \frac{\rho A(\beta)}{\sqrt{n+m}}\sum_{i=1}^{m}\int\left(I(Y_i^{\bar{D}}\le t) - F_{\mathbf{Z^{\bar{D}}}}^{\bar{D}}(t)\right)dF_{\mathbf{Z^D}}^D(t) + o_p(1) \quad (A.4)$$

$$\equiv \frac{1}{\sqrt{n+m}}\sum_{j=1}^{n}\mathbf{H_j} + \frac{1}{\sqrt{n+m}}\sum_{i=1}^{m}\rho A(\beta)B_i + o_p(1) \quad (A.5)$$

$$\equiv \frac{1}{\sqrt{n+m}}\sum_{k=1}^{n+m}W_k(\beta) + o_p(1), \quad (A.6)$$

where $A(\beta) = \lim_n \frac{1}{n}\sum_{j=1}^{n}\omega(\beta^T\mathbf{Z}_j^D)\mathbf{Z}_j^D$, $B_i = \int\left(I(Y_i^{\bar{D}}\le t) - F_{\mathbf{Z^{\bar{D}}}}^{\bar{D}}(t)\right)dF_{\mathbf{Z^D}}^D(t)$, $\rho = \lim_n \frac{n}{m}$, and

$$W_k(\beta) = \begin{cases} \mathbf{H_k}, & \text{if } k = 1,\cdots,n, \\ \rho A(\beta)B_i, & \text{if } k = n+1,\cdots,n+m. \end{cases}$$

Lemma 3.1 follows immediately from (A.1), (A.6) and

$$\frac{1}{\sqrt{n+m}}\sum_{k=1}^{n+m}W_k(\beta) = \sqrt{\frac{n}{n+m}}\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\widehat{\mathbf{H}}_j + o_p(1).$$

Lemma 3.2 *Under the conditions in Theorem 1, we have that*

$$(i) \quad \max_k \parallel \widehat{W}_k(\beta) \parallel = o_p(N^{1/2}),$$

$$(ii) \quad \frac{1}{N}\sum_{j=1}^{N}\widehat{W}_k(\beta)^{\otimes 2} \xrightarrow{p} \frac{\rho}{1+\rho}V,$$

where $N = n+m$.

Proof. (i) Under the conditions in Theorem 1, we have $\max_j \parallel H_j \parallel = o_p(n^{1/2})$. From

$\sup_y \left|\sqrt{n}(\widehat{F}^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(y) - F^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(y))\right| = O_p(1)$, it follows that

$$\begin{aligned}\|\widehat{H}_j - H_j\| &= \|(U_j^D - \widehat{U}_j^D)\omega(\beta_0^T \mathbf{Z}_j^D)\mathbf{Z}_j^D\| \\ &= \left|\widehat{F}^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(Y_j^D) - F^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(Y_j^D)\right| \|\omega({}_0^T \mathbf{Z}_j^D)\mathbf{Z}_j^D\| \\ &\leq \sup_y \left|\widehat{F}^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(y) - F^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(y)\right| \|\omega(\beta_0^T \mathbf{Z}_j^D)\mathbf{Z}_j^D\| = o_p(1),\end{aligned}$$

uniformly for $j = 1, 2, ..., n$. Therefore,

$$\max_j \| \widehat{H}_j \| \leq \max_j \| H_j \| + \max_j \| H_j - H_j \| = o_p(n^{1/2}).$$

Hence,

$$\max_k \| \widehat{W}_k(\beta) \| \leq \max_k \| \widehat{H}_k \| + C \max_k \| \mathbf{Z}_k^D \| = o_p(N^{1/2}),$$

where C is a generic constant.

(ii) From $\sup_y \left|\sqrt{n}(\widehat{F}^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(y) - F^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(y))\right| = O_p(1)$, and $\sup_y \left|\sqrt{n}(\widehat{F}^{D}_{\mathbf{Z}^{\mathbf{D}}}(y) - F^{D}_{\mathbf{Z}^{\mathbf{D}}}(y))\right| = O_p(1)$, we have that

$$\max_k \| \widehat{W}_k(\beta) - W_k(\beta) \| \leq \max_j \| H_j - H_j \| \tag{A.7}$$

$$+C \max_k \left| \int (I(Y_k \leq t) - \widehat{F}^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(t)) d\widehat{F}^{D}_{\mathbf{Z}^{\mathbf{D}}}(t) - \int (I(Y_k \leq t) - F^{\bar{D}}_{\mathbf{Z}^{\bar{\mathbf{D}}}}(t)) dF^{D}_{\mathbf{Z}^{\mathbf{D}}}(t) \right| + o_p(1) \tag{A.8}$$

$$= o_p(1). \tag{A.9}$$

From the proof of Lemma 3.1, it follows that

$$\begin{aligned}\frac{1}{N}\sum_{j=1}^{N} W_k(\beta)^{\otimes 2} &= \frac{1}{N}\sum_{j=1}^{n} H_j^{\otimes 2} + \frac{1}{N}\sum_{i=1}^{m} \rho^2 A(\beta)^{\otimes 2} B_i^2 && \text{(A.10)} \\ &\xrightarrow{p} \frac{\rho}{1+\rho} E(H_1^{\otimes 2}) + \frac{\rho}{1+\rho} \cdot \rho A(\beta)^{\otimes 2} E(B_1^2) && \text{(A.11)} \\ &= \frac{\rho}{1+\rho}(V_1 + V_2) = \frac{\rho}{1+\rho} V. && \text{(A.12)}\end{aligned}$$

Lemma 2 (ii) follows from (A.9) and (A.12) right away.

The Proof of Theorem 3.1

Using Lemma 3.1, Lemma 3.2 (ii) and the similar argument used in Owen (1990), we can prove that

$$\| \nu \| = O_p(N^{-1/2}). \tag{A.13}$$

Then, applying Taylor's expansion, we get that

$$\begin{aligned} l_1(\beta_0) &= 2\sum_{k=1}^{N} \log(1 + \nu^T \widehat{W}_k(\beta)) \\ &= 2\sum_{k=1}^{N} (\nu^T \widehat{W}_k(\beta) - \frac{1}{2}(\nu^T \widehat{W}_k(\beta))^2) + r_{1N} \ , \end{aligned}$$

with

$$|r_{1N}| \leq C \sum_{k=1}^{N} |\nu^T \widehat{W}_k(\beta)|^3 \leq C \| \nu \|^3 \max_j \| \widehat{W}_k(\beta) \| \sum_j \| \widehat{W}_k(\beta) \|^2 = o_p(1) \ .$$

By (3.8), we have

$$\sum_{k=1}^{N} \frac{\widehat{W}_k(\beta)}{1 + \nu^T \widehat{W}_k(\beta)} = \sum_{k=1}^{N} \widehat{W}_k(\beta)[1 - \nu^T \widehat{W}_k(\beta) + \frac{(\nu \widehat{W}_k(\beta))^2}{1 + \nu^T \widehat{W}_k(\beta)}] \tag{A.14}$$

$$= \sum_{k=1}^{N} \widehat{W}_k(\beta) - (\sum_{k=1}^{N} \widehat{\mathbf{H}}_j^{\otimes 2})\nu + \sum_{k=1}^{N} \frac{\widehat{W}_k(\beta)(\nu^T \widehat{W}_k(\beta))^2}{1 + \nu^T \widehat{\mathbf{H}}_j} = 0. \tag{A.15}$$

(A.13), (A.15) and Lemma 3.2 together imply that

$$\nu = (\sum_{k=1}^{N} \widehat{W}_k(\beta)^{\otimes 2})^{-1} \sum_{k=1}^{N} \widehat{W}_k(\beta) + o_p(n^{-1/2}) \ .$$

Again by (3.8), we get that

$$0 = \sum_{k=1}^{N} \frac{\nu^T \widehat{W}_k(\beta)}{1+\nu^T \widehat{\mathbf{H}}_j} = \sum_{k=1}^{N} (\nu^T \widehat{W}_k(\beta))[1 - \nu^T \widehat{W}_k(\beta) + \frac{(\nu^T \widehat{W}_k(\beta))^2}{1+\nu^T \widehat{W}_k(\beta)}] \quad (A.16)$$

$$= \sum_{k=1}^{N} (\nu^T \widehat{W}_k(\beta)) - \sum_{k=1}^{N} (\nu^T \widehat{W}_k(\beta))^2 + \sum_{k=1}^{N} \frac{(\nu^T \widehat{W}_k(\beta))^3}{1+\nu^T \widehat{\mathbf{H}}_j} . \quad (A.17)$$

From (A.13) and Lemma 3.2, we can get

$$\frac{1}{n} \sum_{k=1}^{N} \frac{(\nu^T \widehat{W}_k(\beta))^3}{1+\nu^T \widehat{\mathbf{H}}_j} = o_p(1) ,$$

then we get

$$\sum_{k=1}^{N} \nu \widehat{W}_k(\beta) = \sum_{k=1}^{N} (\nu \widehat{\mathbf{H}}_j)^2 + o_p(1) . \quad (A.18)$$

From $(A.13) - (A.18)$, Lemma 3.1, and Lemma 3.2 (ii), it follows that

$$\begin{aligned} l_1(\beta_0) &= \sum_{k=1}^{N} \nu^T \widehat{W}_k(\beta)^{\otimes 2} \nu + o_p(1) \\ &= (\frac{1}{\sqrt{N}} \sum_{k=1}^{N} \widehat{W}_k(\beta))^T (\frac{1}{N} \sum_{k=1}^{N} \widehat{W}_k(\beta)^{\otimes 2})^{-1} (\frac{1}{\sqrt{N}} \sum_{k=1}^{N} \widehat{W}_k(\beta)) + o_p(1) \\ &\xrightarrow{\mathcal{L}} \chi^2_p. \end{aligned}$$

The Proof of Theorem 3.2.

The proof of Theorem 3.2 is similar to that of the main theorem in Jing, Yuan, and Zhou (2009) and hence omitted here.

Appendix B

PROOFS OF PART 4

Lemma 4.1

(i). $n^{-1/2}\sum_{i=1}^{n}\widehat{V}(W_i)-\rho \xrightarrow{\mathcal{L}} N(0,\sigma_V^2)$,

where $\sigma_V^2 = Var[\frac{X-\mu_x}{\sigma_x}\cdot\frac{Y_i-\mu_y}{\sigma_y} - 2^{-1}\rho((\frac{X-\mu_x}{\sigma_x})^2 + (\frac{Y-\mu_y}{\sigma_y}{}^2)]$.

(ii). $\frac{1}{n}\sum_{i=1}^{n}(\widehat{V}(W_i)-\rho)^2 \xrightarrow{p} \sigma_0^2$, where $\sigma_0^2 = Var[\frac{X-\mu_x}{\sigma_x}\cdot\frac{Y_i-\mu_y}{\sigma_y}]$.

Proof of Lemma 4.1

(i). Let $V(W_i) = \frac{X_i-\mu_x}{S_X}\cdot\frac{Y_i-\mu_y}{S_Y},\ i=1,\ldots,n$. Then, we have the following decomposition:

$$n^{-1/2}\sum_{i=1}^{n}(\widehat{V}(W_i)-\rho) = n^{-1/2}\sum_{i=1}^{n}(\widehat{V}(W_i)-V(W_i)) + n^{-1/2}\sum_{i=1}^{n}(V(W_i)-V(W_i)) \quad \text{(B.1)}$$

$$+n^{-1/2}\sum_{i=1}^{n}(V(W_i)-\rho) \equiv I_1+I_2+I_3. \quad \text{(B.2)}$$

For the firm term in (B.2), we have that

$$I_1 = n^{-1/2}\sum_{i=1}^{n}(\widehat{V}(W_i)-V(W_i)) \quad \text{(B.3)}$$

$$= (S_XS_Y)^{-1}n^{-1/2}\sum_{i=1}^{n}[(X_i-\bar{X})(Y_i-\bar{Y})-(X_i-\mu_x)(Y_i-\mu_y)] \quad \text{(B.4)}$$

$$= (S_XS_Y)^{-1}n^{-1/2}[(\mu_x-\bar{X})\sum_{i=1}^{n}(y_i-\bar{Y})+(\mu_y-\bar{Y})\sum_{i=1}^{n}(X_i-\mu_x)] = o_p(1). \quad \text{(B.5)}$$

For the firm term in (B.2), we have that

$$\begin{aligned} I_1 &= n^{-1/2}\sum_{i=1}^{n}(\widehat{V}(W_i) - V(W_i)) && \text{(B.6)} \\ &= (S_X S_Y)^{-1} n^{-1/2}\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y}) - (X_i - \mu_x)(Y_i - \mu_y)] && \text{(B.7)} \\ &= (S_X S_Y)^{-1} n^{-1/2}[(\mu_x - \bar{X})\sum_{i=1}^{n}(y_i - \bar{Y}) + (\mu_y - \bar{Y})\sum_{i=1}^{n}(X_i - \mu_x)] = o_p(1). && \text{(B.8)} \end{aligned}$$

For the second term in (B.2), we have that

$$\begin{aligned} I_2 &= n^{-1/2}\sum_{i=1}^{n}(V(W_i) - V(W_i)) && \text{(B.9)} \\ &= (\frac{1}{S_X S_Y} - \frac{1}{\sigma_x \sigma_y}) n^{-1/2}\sum_{i=1}^{n}(X_i - \mu_x)(Y_i - \mu_y) && \text{(B.10)} \\ &= -\rho(S_X S_Y)^{-1} n^{1/2}[(S_X - \sigma_x)S_Y + (S_Y - \sigma_y)\sigma_x] + o_p(1) && \text{(B.11)} \\ &= -2^{-1}\rho n^{1/2}[\sigma_x^{-2}(s_x^2 - \sigma_x^2) + \sigma_y^{-2}(S_Y^2 - \sigma_y^2)] + o_p(1) && \text{(B.12)} \\ &= -2^{-1}\rho n^{-1/2}\sum_{i=1}^{n}[((\frac{X_i - \mu_x}{\sigma_x})^2 - 1) + ((\frac{Y_i - \mu_y}{\sigma_y})^2 - 1)] + o_p(1). && \text{(B.13)} \end{aligned}$$

As for the third term in (B.2), we have that

$$\begin{aligned} I_3 &= n^{-1/2}\sum_{i=1}^{n}(V(W_i) - \rho) && \text{(B.14)} \\ &= n^{-1/2}\sum_{i=1}^{n}[\frac{X_i - \mu_x}{\sigma_x} \cdot \frac{Y_i - \mu_y}{\sigma_y} - \rho]. && \text{(B.15)} \end{aligned}$$

From (B.2) - (B.15) and the central limit theorem, it follows that

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n}(\widehat{V}(W_i) - \rho) &= n^{-1/2}\sum_{i=1}^{n}[(\frac{X_i - \mu_x}{\sigma_x} \cdot \frac{Y_i - \mu_y}{\sigma_y} - \rho) \\ &\quad -2^{-1}\rho(((\frac{X_i - \mu_x}{\sigma_x})^2 - 1) + ((\frac{Y_i - \mu_y}{\sigma_y})^2 - 1))] + o_p(1) \\ &\xrightarrow{\mathcal{L}} N(0, \sigma_V^2). \end{aligned}$$

Lemma (ii) follows directly from the law of large number and

$$\frac{1}{n}\sum_{i=1}^{n}(V(W_i)-\rho)^2 \xrightarrow{p} \sigma_0^2,$$

as well as

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}(\widehat{V}(W_i)-\rho)^2-\frac{1}{n}\sum_{i=1}^{n}(V(W_i)-\rho)^2\right| \le O_p(1)\frac{1}{n}\sum_{i=1}^{n}|\widehat{V}(W_i)-V(W_i)| \\
&\le O_p(1)\frac{1}{n}[|\mu_y-\bar{Y}|\sum_{i=1}^{n}|X_i|+|\mu_x-\bar{X}|\sum_{i=1}^{n}|Y_i|+\sum_{i=1}^{n}|\bar{X}\bar{Y}-\mu_x\mu_y|]+o_p(1) \\
&= o_p(1).
\end{aligned}$$

Proof of Theorem 4.1.

Using Lemma 1 and the similar arguments used in Owen (1990), we can prove that $\lambda = O_p(n^{-1/2})$. Since $max_{1\le i\le n}|V(W_i)-\rho| = O(1)$, a.s., applying Taylor's expansion to (4.8), we obtain that

$$\begin{aligned}
l(\rho) &= 2\sum_{i=1}^{n}\log[1+\lambda(\widehat{V}(W_i)-\rho)] \\
&= 2\sum_{i=1}^{n}[\lambda(\widehat{V}(W_i)-\rho)-\frac{1}{2}(\lambda(\widehat{V}(W_i)-\rho))^2]+r_n,
\end{aligned}$$

where

$$|r_n| \le C\sum_{i=1}^{n}|\lambda(\widehat{V}(W_i)-\rho)|^3 \le C|\lambda|^3 n = O_p(n^{-1/2}).$$

From equations (4.6), it follows that

$$\lambda = \frac{\sum_{i=1}^{n}(\widehat{V}(W_i)-\rho)}{\sum_{i=1}^{n}(\widehat{V}(W_i)-\rho)^2}+O_p(n^{-1/2}),$$

$$\sum_{i=1}^{n} \lambda(\widehat{V}(W_i) - \rho) = \sum_{i=1}^{n} (\lambda(V(W_i) - \rho))^2 + o_p(1).$$

Therefore, by Lemma 1, we have that

$$\begin{aligned} A \cdot l(\rho) &= A \cdot \sum_{i=1}^{n} (\widehat{V}(W_i) - \rho) + o_p(1) \\ &= \frac{\sigma_0^2}{\sigma_V^2} \cdot \frac{[\sum_{i=1}^{n} (\widehat{V}(W_i) - \rho)]^2}{\sum_{i=1}^{n} (\widehat{V}(W_i) - \rho)^2} + o_p(1) \\ &= \frac{\sigma_0^2}{n^{-1} \sum_{i=1}^{n} (\widehat{V}(W_i) - \rho)^2} \cdot [\sigma_V^{-1} n^{-1/2} \sum_{i=1}^{n} (\widehat{V}(W_i) - \rho)]^2 + o_p(1) \\ &\xrightarrow{\mathcal{L}} \chi_1^2. \end{aligned}$$

The proof of Theorem 1 is thus completed.

Proof of Theorem 4.2.

Proof of Theorem 4.2 is similar to that of Theorem 1 and hence omitted here.