SCIENCE CHINA
Mathematics

Methods for Diagnostic Tests and Biomarkers

August 2012 Vol. 55 No. 8: 1553–1564
doi: 10.1007/s11425-012-4455-2

• ARTICLES •

Empirical likelihood-based inferences for the area under the ROC curve with covariates

YANG BaoYing[1] & QIN GengSheng[2,*]

[1]*College of Mathematics, Southwest Jiaotong University, Chengdu* 610031*, China;*
[2]*Department of Mathematics and Statistics, Georgia State University, Atlanta, GA* 30303*, USA*
Email: yangby926@yahoo.com.cn, gqin@gsu.edu

Received May 14, 2012; accepted May 17, 2012; published online July 12, 2012

Abstract In the receiver operating characteristic (ROC) analysis, the area under the ROC curve (AUC) is a popular summary index of discriminatory accuracy of a diagnostic test. Incorporating covariates into ROC analysis can improve the diagnostic accuracy of the test. Regression model for the AUC is a tool to evaluate the effects of the covariates on the diagnostic accuracy. In this paper, empirical likelihood (EL) method is proposed for the AUC regression model. For the regression parameter vector, it can be shown that the asymptotic distribution of its EL ratio statistic is a weighted sum of independent chi-square distributions. Confidence regions are constructed for the parameter vector based on the newly developed empirical likelihood theorem, as well as for the covariate-specific AUC. Simulation studies were conducted to compare the relative performance of the proposed EL-based methods with the existing method in AUC regression. Finally, the proposed methods are illustrated with a real data set.

Keywords AUC regression, bootstrap, confidence region, empirical likelihood

MSC(2010) 62G15, 62G20, 62P20

Citation: Yang B Y, Qin G S. Empirical likelihood-based inferences for the area under the ROC curve with covariates. Sci China Math, 2012, 55(8): 1553–1564, doi: 10.1007/s11425-012-4455-2

1 Introduction

Diagnostic test is one of the most important components in modern medical studies. Evaluating the accuracy of a diagnostic test is indispensable in medical diagnostics [16]. For a continuous-scale diagnostic test, we assume that an individual is classified as diseased if the measurement of the test exceeds a chosen threshold value, and otherwise classified as non-diseased. The *Receiver Operating Characteristic* (ROC) curve of the diagnostic test is the plot of sensitivity versus one minus specificity for all possible threshold values (see [11, 19]). The area under the ROC curve (AUC) is the summary index of the discriminatory accuracy of an ROC curve. Let Y^D be the response of a diseased individual and $Y^{\bar{D}}$ be the response of a non-diseased individual. The AUC can be expressed as $\mathrm{AUC} = P(Y^D > Y^{\bar{D}})$ (see [1]). The closer to one the AUC value, the more accurate the diagnostic test.

When performing a diagnostic test, some factors like characteristics of study subjects or operating conditions for the test may affect the results by influencing the distributions of test measurements for "diseased" and/or "non-diseased" subjects. Pepe [11] gave a wonderful introduction to why and how to adjust for covariates in the ROC curve. When covariates are available, regression analysis is a powerful and useful tool. Tosteson and Begg [18] modeled the test result as a function of diseased status and

*Corresponding author

covariates. But their method cannot be used to compare different types of tests (see [12]). Thompson and Zucchini [17] and Obuchowski [6] proposed AUC regression methods based on the derived variable. Dorfman, Berbaum, and Metz [3] developed a method based on the computing jackknifed AUC values for each subject. However, Dodd and Pepe [2] pointed out that these methods can only accommodate discrete covariates, they proposed a regression model for the AUC summary measure. Lin et al. [5] studied the ROC regression model with completely unknown link and baseline functions, and proposed a powerful semiparametric procedure to estimate both the parametric and nonparametric components of the model.

Assume that there are a sample of m non-diseased subjects with responses $Y_i^{\bar{D}}$ and covariate vectors $\mathbf{Z}_i^{\bar{D}} = (Z_{i1}^{\bar{D}}, \ldots, Z_{iq}^{\bar{D}})^{\mathrm{T}},\ i = 1, 2, \ldots, m$, and a sample of n diseased subjects with responses Y_j^D and covariate vectors $\mathbf{Z}_j^D = (Z_{j1}^D, \ldots, Z_{jp}^D)^{\mathrm{T}},\ j = 1, 2, \ldots, n$. Suppose that $Y^{\bar{D}}$ follows the distribution function $F^{\bar{D}}_{\mathbf{Z^{\bar{D}}}}$ and Y^D follows the distribution function $F^D_{\mathbf{Z^D}}$. 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 [2] defined the AUC regression model as follows:

$$\theta_{ij} = g(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{Z}_{ij}),$$

where $\mathbf{Z}_{ij}$ denotes the observable covariates $(\mathbf{Z}_i^{\bar{D}}, \mathbf{Z}_j^D)$, and g is a specified function. They proposed the following generalized estimating equation to obtain the estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$:

$$\sum_{j=1}^{n_D} \sum_{i=1}^{n_{\bar{D}}} \frac{\partial \theta_{ij}}{\partial \boldsymbol{\beta}} \omega(\mathbf{Z}_{ij}, \boldsymbol{\beta})(I_{ij} - g(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{Z}_{ij})) = 0, \tag{1.1}$$

where $I_{ij} = I(Y_j^D > Y_i^{\bar{D}})$, and $\omega(\mathbf{Z}_{ij}, \boldsymbol{\beta})$ is a known weight function. Furthermore, they proved that

$$\sqrt{\frac{n_{\bar{D}} n_D}{N}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} N(0, \Xi), \quad \text{as} \quad N \to \infty, \tag{1.2}$$

where $N = n_D + n_{\bar{D}}$, and Ξ is the asymptotic variance of $\widehat{\boldsymbol{\beta}}$.

The asymptotic variance Ξ in (1.2) is still unknown and of a very complex form. This asymptotic normal distribution may be used to construct confidence regions/intervals for the parameter vector $\boldsymbol{\beta}$ if a good estimate for Ξ can be obtained. However, no explicit estimate for Ξ was proposed in Dodd and Pepe [2]. They proposed to use bootstrap method to estimate Ξ. Yet, our simulation studies indicated that the normal approximation-based confidence regions for $\boldsymbol{\beta}$ may have poor coverage accuracy by using the bootstrap variance estimate.

Empirical likelihood (EL), introduced by Owen [8, 9], is a powerful non-parametric method and its advantages over the normal approximation-based methods have been well-recognized (e.g., [4]). Over last two decades, empirical likelihood has found wide applications in many areas such as in econometrics, medical studies and survey sampling. Readers are referred to Owen [7] and references therein. Qin and Zhou [14] successfully applied empirical likelihood to inferences on the AUC. They constructed an EL-based confidence interval for the AUC using the scaled chi-square distribution, and the EL-based interval has better finite sample performances than the existing bootstrap intervals and normal approximation-based intervals. Their work inspires us to develop new EL-based inferences on the AUC with covariates. In this paper, we will develop a hybrid bootstrap and empirical likelihood method for the AUC regression model.

The paper is organized as follows. In Section 2, we define the profile EL ratio statistic for the parameter vector in the AUC regression model. The asymptotic distribution of the EL ratio statistic is shown to be a weighted sum of independent chi-square distributions. In Section 3, we propose EL-based confidence regions for the parameter vector and a confidence interval for the covariate-specific AUC. In Section 4, we report the results of the simulation studies that compare the proposed EL-based methods with the existing method. In Section 5, we apply the recommended methods to an audiology dataset. Finally, the proof of the main theorem is presented in Appendix.

2 EL for the AUC regression model

Let Y be the continuous-scale test result which is sought to discriminate between the "diseased" and "non-diseased" populations. The test result Y can be standardized according to the distribution in the non-diseased population $F^{\bar{D}}_{\mathbf{Z}^{\bar{D}}}$, so called the *reference distribution* in Pepe and Cai [10]. The *placement value* of Y is defined as $U = 1 - F^{\bar{D}}_{\mathbf{Z}^{\bar{D}}}(Y)$. In particular, $U^{\bar{D}} = 1 - F^{\bar{D}}_{\mathbf{Z}^{\bar{D}}}(Y^{\bar{D}})$ is uniform in $(0,1)$, and $U^D = 1 - F^{\bar{D}}_{\mathbf{Z}^{\bar{D}}}(Y^D)$ measures the separation between the diseased and non-diseased groups. The ROC curve and the placement value have closed relationship. If setting $u = 1 - F^{\bar{D}}_{\mathbf{Z}^{\bar{D}}}(y)$ as a false positive rate, then the conditional distribution of the placement value U^D given $\mathbf{Z}^D$ is the conditional ROC curve. i.e., $\mathrm{ROC}_{\mathbf{Z}}(u) = P\ (U^D < u|\mathbf{Z}^D)$. The covariate-specific AUC can be expressed as $\mathrm{AUC}_{\mathbf{Z}} = E(1 - U^D|\mathbf{Z}^D)$. To evaluate the covariate effects on discrimination, we use the following AUC regression model:

$$\mathrm{AUC}_{\mathbf{Z}} = E(1 - U^D|\mathbf{Z}^{\mathbf{D}}) = g(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}^D), \tag{2.1}$$

where $\mathbf{Z}^D = (Z_1^D, \ldots, Z_p^D)^{\mathrm{T}}$ is the $p \times 1$ vector of covariates, g is a specified link function. To estimate $\boldsymbol{\beta}$, we use the weighted least squares fitting-based estimation equation:

$$\sum_{j=1}^{n} \omega(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j^D)\left[1 - U_j^D - g(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j^D)\right]\mathbf{Z}_j^D = \mathbf{0}, \tag{2.2}$$

where $\omega(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j^D)$ is a given scalar weight function, $\mathbf{0} = (0, \ldots, 0)^{\mathrm{T}}$ 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.

Based on estimation equation (2.2), the empirical likelihood for $\boldsymbol{\beta}$ can be defined as

$$\tilde{L}(\boldsymbol{\beta}) = \sup\left\{\prod_{j=1}^{n} p_j : \sum_{j=1}^{n} p_j = 1, \sum_{j=1}^{n} p_j(1 - U_j^D - g(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j^D))\omega(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D = \mathbf{0}\right\} \tag{2.3}$$

where p_j denotes the probability mass at U_j^D, $j = 1, 2, \ldots, n$. Here, the placement value $U^D = 1 - F^{\bar{D}}_{\mathbf{Z}^{\bar{D}}}(Y^D)$ is unobservable because the distribution for the test result in the non-diseased population is unknown. However, using the non-diseased sample $\{(Y_i^{\bar{D}}, \mathbf{Z}_i^{\bar{D}}), i = 1, 2, \ldots, m\}$, we can estimate the reference distribution $F^{\bar{D}}_{\mathbf{Z}^{\bar{D}}}$ by its empirical distribution $\widehat{F}^{\bar{D}}_{\mathbf{Z}^{\bar{D}}}$. Therefore, the unknown placement value U_j^D can be replaced by the estimated placement value $\widehat{U}_j^D = 1 - \widehat{F}^{\bar{D}}_{\mathbf{Z}^{\bar{D}}}(Y_j^D)$. Let

$$\mathbf{H}_j = (1 - U_j^D - g(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j^D))\omega(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D,$$
$$\widehat{\mathbf{H}}_j = (1 - \widehat{U}_j^D - g(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j^D))\omega(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D.$$

Then, the profile empirical likelihood for $\boldsymbol{\beta}$ can be defined as

$$L(\boldsymbol{\beta}) = \sup\left\{\prod_{j=1}^{n} p_j : \sum_{j=1}^{n} p_j = 1, \sum_{j=1}^{n} p_j\widehat{\mathbf{H}}_j = \mathbf{0}\right\}.$$

Using the Lagrange multiplier method, we have

$$p_j = \frac{1}{n}\frac{1}{1 + \boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j},$$

where $\boldsymbol{\nu}^{\mathrm{T}} = (\nu_1, \nu_2, \ldots, \nu_p)$ is the solution to

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\widehat{\mathbf{H}}_j}{1 + \boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j} = \mathbf{0}. \tag{2.4}$$

Note that $\prod_{j=1}^{n} p_i$, subject to $\sum_{j=1}^{n} p_j = 1, p_j \geqslant 0, j = 1, 2, \ldots, n$, attains its maximum n^{-n} at $p_j = n^{-1}$. So, the profile empirical likelihood ratio for $\boldsymbol{\beta}$ is

$$R(\boldsymbol{\beta}) = \prod_{j=1}^{n}(np_j) = \prod_{j=1}^{n}\{1 + \boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j\}^{-1}.$$

The corresponding profile empirical log-likelihood ratio for $\boldsymbol{\beta}$ is

$$l(\boldsymbol{\beta}) = -2\log R(\boldsymbol{\beta}) = 2\sum_{j=1}^{n}\log(1+\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j).$$

Theorem 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, the matrix* $\mathbf{G} = \mathbf{V}_0^{-1}\mathbf{V}$, *which will be defined in the Appendix, is positively definite, and* $\boldsymbol{\beta}_0$ *is the true parameter vector in the* AUC *regression model, then the asymptotic distribution of the profile empirical log-likelihood ratio* $l(\boldsymbol{\beta}_0)$ *is a weighted sum of independent chi-square distributions with one degree of freedom. That is,*

$$l(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} k_1\chi_{1,1}^2 + \cdots + k_p\chi_{p,1}^2, \tag{2.5}$$

where $\chi_{i,1}^2$, $i = 1, 2, \ldots, p$, *are independent chi-square random variables with one degree of freedom, and the weights* k_i*'s are the eigenvalues of* $\mathbf{G}$.

Note that the empirical log-likelihood ratio $l(\boldsymbol{\beta})$ is not a sum of independent random variables. By Theorem 1, its asymptotic distribution is a weighted sum of independent chi-square distributions rather than the standard chi-square distribution.

3 EL-based inferences in the AUC regression model

In this section, we will propose EL-based confidence regions for the regression parameter vector $\boldsymbol{\beta}$ and a confidence interval for the covariate-specific AUC.

3.1 EL-based confidence region for $\boldsymbol{\beta}$

The empirical likelihood theory developed in Section 2 can be used to construct confidence region for $\boldsymbol{\beta}$. Let c_α be the $(1-\alpha)$-th quantile of the asymptotic weighted chi-square distribution in Theorem 1. Then

$$\mathrm{CR}(\boldsymbol{\beta}) = \{\boldsymbol{\beta} : l(\boldsymbol{\beta}) \leqslant c_\alpha\}$$

is a $(1-\alpha)$-th confidence region for the regression parameter vector $\boldsymbol{\beta}$. That is,

$$P\{\boldsymbol{\beta}_0 \in \mathrm{CR}(\boldsymbol{\beta})\} = P\{l(\boldsymbol{\beta}) \leqslant c_\alpha\} = 1-\alpha+o(1).$$

By Theorem 1, the calculation of c_α involves in the estimation of matrix $\mathbf{G}$. Here we propose a *Hybrid Bootstrap and Empirical Likelihood* (*HBEL*) approach to estimate c_α without estimating $\mathbf{G}$. We summarize the procedure in the following steps:

1. Generate bootstrap re-samples $\{Y_i^{*\bar{D}}, \mathbf{Z}_i^{*\bar{D}} : i = 1, 2, \ldots, m\}$ and $\{Y_j^{*D}, \mathbf{Z}_j^{*D} : j = 1, 2, \ldots, n\}$ from the original samples $\{Y_i^{\bar{D}}, \mathbf{Z}_i^{\bar{D}} : i = 1, 2, \ldots, m\}$ and $\{Y_j^D, \mathbf{Z}_j^D : j = 1, 2, \ldots, n\}$, respectively.

2. Calculate the bootstrap versions $F_{\mathbf{Z}^{\bar{D}}}^{*\bar{D}}$, $\widehat{U}_j^{*D}$, $\widehat{\boldsymbol{\beta}}^*$ and $\widehat{\mathbf{H}}_j^*$ of $F_{\mathbf{Z}^{\bar{D}}}^{\bar{D}}$, $\widehat{U}_j^D$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{H}}_j$, where

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

3. Find $\boldsymbol{\nu}^*$ by solving

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\widehat{\mathbf{H}}_j^*}{1+\boldsymbol{\nu}^{*\mathrm{T}}\widehat{\mathbf{H}}_j^*} = 0.$$

Then compute the bootstrap version $l^*(\widehat{\boldsymbol{\beta}}^*)$ of $l(\boldsymbol{\beta}_0)$:

$$l^*(\widehat{\boldsymbol{\beta}}^*) = 2\sum_{j=1}^{n}\log(1+\boldsymbol{\nu}^{*\mathrm{T}}\widehat{\mathbf{H}}_j^*).$$

4. Repeat Steps 1–3 B (it is recommended that $B \geqslant 200$; in this paper, we take $B = 500$) times to get B bootstrap replications: $\{l^*_{1,b}(\widehat{\boldsymbol{\beta}}^*) : b = 1, 2, \ldots, B\}$.

We propose two $(1-\alpha)$-th HBEL confidence regions for $\boldsymbol{\beta}$ as follows:

(i) HBEL1 confidence region: $\mathrm{CR}_1(\boldsymbol{\beta}) = \{\boldsymbol{\beta} : l(\boldsymbol{\beta}) \leqslant l^*_{([B(1-\alpha)])}(\widehat{\boldsymbol{\beta}}^*)\}$, where $l^*_{([B(1-\alpha)])}(\widehat{\boldsymbol{\beta}}^*)$ is the $[B(1-\alpha)]$-th ordered value of $l^*_{1,b}(\widehat{\boldsymbol{\beta}}^*)$'s.

(ii) HBEL2 confidence region : $\mathrm{CR}_2(\boldsymbol{\beta}) = \{\boldsymbol{\beta} : l(\boldsymbol{\beta}) \leqslant a\chi^2_{p,1-\alpha}\}$, where $a = \frac{1}{pB}\sum_{b=1}^{B} l^*_{1,b}(\widehat{\boldsymbol{\beta}}^*)$.

3.2 EL-based confidence interval for $\mathrm{AUC}_{\mathbf{Z}}$

In practice, we may need a confidence interval for a covariate-specific AUC. Let $\mathrm{CR}(\boldsymbol{\beta})$ be a $(1-\alpha)$-th confidence region for $\boldsymbol{\beta}$. For a specified covariate $\mathbf{Z}^D$, the confidence interval for the covariate-specific AUC can be constructed as follows:

$$\{\mathrm{AUC}_Z(\boldsymbol{\beta}) = g(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}^D) : \boldsymbol{\beta} \in \mathrm{CR}(\boldsymbol{\beta})\}. \tag{3.1}$$

This interval is a $(1-\alpha)$-th confidence interval for the covariate-specific AUC when $g(\cdot)$ is a one-to-one function.

We propose the following procedure to compute the interval (see also [20]). Note that the confidence interval for AUC_Z can be written as (q_0, q_1) for large N,

$$\begin{aligned} q_0 &= \min\{\mathrm{AUC}_Z(\boldsymbol{\beta}) : \boldsymbol{\beta} \in CR(\boldsymbol{\beta})\} = \min\{\mathrm{AUC}_Z(\boldsymbol{\beta}) : l(\boldsymbol{\beta}) = c, 0 \leqslant c \leqslant c_\alpha\} \\ &\approx \min\left\{\bigcup_{i=1}^{N} (\mathrm{AUC}_Z(\boldsymbol{\beta}) : l(\boldsymbol{\beta}) = c_i)\right\}, \\ q_1 &= \max\{\mathrm{AUC}_Z(\boldsymbol{\beta}) : \boldsymbol{\beta} \in CR(\boldsymbol{\beta})\} = \max\{\mathrm{AUC}_Z(\boldsymbol{\beta}) : l(\boldsymbol{\beta}) = c, 0 \leqslant c \leqslant c_\alpha\} \\ &\approx \max\left\{\bigcup_{i=1}^{N} (\mathrm{AUC}_Z(\boldsymbol{\beta}) : l(\boldsymbol{\beta}) = c_i)\right\}, \end{aligned}$$

where $c_1, c_2, \ldots, c_N$ is a random sample of size N generated from the uniform on $[0, c_\alpha]$.

To estimate q_0, q_1, first we approximate $\mathrm{CR}(\boldsymbol{\beta})$ by $\mathrm{CR}_0 = \{\boldsymbol{\beta} : \widehat{\boldsymbol{\beta}}_k - z_{1-\alpha/2}\widehat{\boldsymbol{\sigma}}_k \leqslant \boldsymbol{\beta}_k \leqslant \widehat{\boldsymbol{\beta}}_k + z_{1-\alpha/2}\widehat{\boldsymbol{\sigma}}_k,\ k = 1, 2, \ldots, p\}$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th quantile of the standard normal distribution, and $\widehat{\boldsymbol{\sigma}}_k$ is the standard error of $\widehat{\boldsymbol{\beta}}_k$. By generating J (J is an appropriate chosen integer depending on the number of regression parameters) vectors $\{\boldsymbol{\beta}^{(j)} :\ j = 1, 2, \ldots, J\}$ uniformly over CR_0 that are satisfying $l(\boldsymbol{\beta}^{(j)}) \leqslant c_\alpha$, we can estimate $\boldsymbol{\beta}$ to satisfy $l(\boldsymbol{\beta}) = c,\ \forall\, c \in [0, c_\alpha]$ by smoothing technique (for example, local linear method) based on the data $\{(\boldsymbol{\beta}^{(j)}, l(\boldsymbol{\beta}^{(j)}) :\ j = 1, 2, \ldots, J\}$.

4 A simulation study

In this section, we conduct a simulation study to compare the EL-based confidence regions for $\boldsymbol{\beta}$ with the existing normal approximation-based confidence region in terms of coverage probability.

Normal approximation is a commonly used method for constructing confidence region for a regression parameter vector. As mentioned in Section 1, Dodd and Pepe [2] proposed an estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ and obtained the asymptotic normal distribution of the estimator. So, a normal approximation-based confidence region (NA) for $\boldsymbol{\beta}$ can be constructed as follows:

$$\mathrm{CR}_{NA}(\boldsymbol{\beta}) = \left\{\boldsymbol{\beta} : \frac{N}{n_{\bar{D}}n_D} \cdot (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}}\Xi^{*-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \leqslant \chi^2_{p,1-\alpha}\right\}, \tag{4.1}$$

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

In the study, we use the similar simulation setting as in Dodd and Pepe [2]. Let $Y^D|\mathbf{Z}^D \sim N(\mu_{D,\mathbf{Z}^D}, \sigma^2_D)$ and $Y^{\bar{D}}|\mathbf{Z}^{\bar{D}} \sim N(\mu_{\bar{D},\mathbf{Z}^{\bar{D}}}, \sigma^2_{\bar{D}})$. Then, we have

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

Two AUC regression models with dimension $p = 2, 3$ are considered in the study. In both models, the covariates $\mathbf{Z}^D$ and $\mathbf{Z}^{\bar{D}}$ are assumed to have common components. In the first model with $p = 2$, the AUC regression model has only one continuous covariate Z with $Z \sim U(0, 10)$, and $\mu_{\bar{D},Z} = \gamma_0 + \gamma_1 Z$, $\mu_{D,Z} = k_0 + k_1 Z$. Then

$$\mathrm{AUC}_Z = \Phi\left(\frac{(k_0 - \gamma_0) + (k_1 - \gamma_1)Z}{\sqrt{\sigma_D^2 + \sigma_{\bar{D}}^2}}\right) = \Phi\left(\beta_0 + \beta_1 Z\right), \tag{4.2}$$

where $\beta_0 = (k_0 - \gamma_0)/\sqrt{\sigma_D^2 + \sigma_{\bar{D}}^2}$, $\beta_1 = (k_1 - \gamma_1)/\sqrt{\sigma_D^2 + \sigma_{\bar{D}}^2}$. We choose $\gamma_0 = 0$, $\gamma_1 = 0$, $k_0 = 0$, $k_1 = 0.5$, $\sigma_D = 1.2$, $\sigma_{\bar{D}} = 1$ so that the true parameters $\beta_0 = 0$, $\beta_1 = 0.32$. The sample size m, n are chose to be 30, 50, 100 respectively in both diseased group and non-diseased group. After generating 1000 random samples based on the above simulation setting and using the usual generalized linear regression method, we can obtain estimate $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ and calculate the coverage probabilities of the NA-based confidence region and the proposed HBEL confidence regions for $\boldsymbol{\beta}$. The parameter estimates and the coverage probabilities for $\boldsymbol{\beta}$ are presented in Tables 1 and 2.

From Table 1, we can see that the GLM method provides good estimates for $\boldsymbol{\beta}$. From Table 2, we observe that the coverage probabilities of HBEL1 confidence region are very close to the nominal confidence levels. It performs the best among three confidence regions for $\boldsymbol{\beta}$. When sample size is small $(m, n = 30)$, the NA confidence region over-covers the true regression parameters. This is also mentioned in Dodd and Pepe [2]. When sample size is large (e.g., $m, n = 100$), the NA confidence region has acceptable coverage probability.

In the second model with $p = 3$, the AUC regression model has two continuous covariates Z_1 and Z_2 with $Z_1 \sim U(0, 10)$ and $Z_2 \sim N(1.2, 3)$. Assume that $\mu_{\bar{D},Z} = \gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2$, and $\mu_{D,Z} = k_0 + k_1 Z_1 + k_2 Z_2$. Similar to (4.2), we have that $\mathrm{AUC}_{\mathbf{Z}} = \Phi\left(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2\right)$. We choose $\gamma_0 = 0.7$, $\gamma_1 = -0.3$, $\gamma_2 = 0.3$, $k_0 = 1$, $k_1 = -0.8$, $k_2 = 1.7$, $\sigma_D = 1.2$, $\sigma_{\bar{D}} = 1$ so that the true parameters $\beta_0 = 0.192$, $\beta_1 = -0.32$, and $\beta_2 = 0.896$. Based on this simulation setting, we do similar computation to

Table 1 The parameters estimates in the AUC regression model $\mathrm{AUC}_{\mathbf{Z}} = \Phi(\beta_0 + \beta_1 Z_1)$

n	m	β_0	$\hat{\beta}_0$	Bias of $\hat{\beta}_0$	sd of $\hat{\beta}_0$
30	30	0	−0.03247	−0.03247	0.19512
50	50	0	−0.02364	−0.02364	0.08235
100	100	0	−0.01362	−0.01362	0.04350
n	m	β_1	$\hat{\beta}_1$	Bias of $\hat{\beta}_1$	sd of $\hat{\beta}_1$
30	30	0.3201	0.35335	0.03326	0.01788
50	50	0.3201	0.33580	0.01571	0.00624
100	100	0.3201	0.32951	0.00942	0.00307

Table 2 Coverage probabilities of 90% and 95% confidence regions for the parameters vector in the AUC regression model $\mathrm{AUC}_{\mathbf{Z}} = \Phi(\beta_0 + \beta_1 Z_1)$

Level	n	m	HBEL1	HBEL2	NA
90%	30	30	0.905	0.888	0.958
	50	50	0.907	0.887	0.927
	100	100	0.895	0.906	0.893
95%	30	30	0.954	0.933	0.98
	50	50	0.948	0.922	0.959
	100	100	0.949	0.931	0.941

obtain parametric estimates for $\boldsymbol{\beta}$ and coverage probabilities of confidence regions for $\boldsymbol{\beta}$. The results are presented in Tables 3 and 4.

From Table 3, we observe that the GLM method provides acceptable estimates for β_1 and β_2. But the estimates for β_0 have sizable biases and standard errors. One possible reason for the biased estimation on β_0 is that the true placement value U^D is unobservable. To estimate U^D, we have to estimate the distribution function $F^{\bar{D}}_{Z^{\bar{D}}}$ of the non-diseased population. In the simulation study, the empirical distribution function was used to estimate $F^{\bar{D}}_{Z^{\bar{D}}}$; more accurate estimate like kernel distribution estimate, which involves in smoothing parameter selection, may be needed for the estimation. Although the estimate for the intercept term β_0 is biased, the regression parameters estimates which reflect the effects of covariates on the AUC are still acceptable.

Table 4 indicates that the coverage probabilities of the NA confidence regions are far below the nominal confidence levels. The poor performances of the NA method is possibly due to the poor estimates for β_0 and the asymptotic variance of $\widehat{\boldsymbol{\beta}}$. However, the HBEL1 and HBEL2 confidence regions have much better coverage accuracy than the NA confidence regions although $\hat{\beta}_0$ is biased. They are more robust than the NA-based method because the EL-based confidence regions have data-determined shapes, and the NA-based region is a symmetric confidence region which is sensitive to the estimates for β_0 and the asymptotic variance of $\widehat{\boldsymbol{\beta}}$. Once again, we observe that the HBEL1 confidence region outperforms the other confidence regions.

5 An illustration example

As an illustration, we analyze the audiology data reported by Stover et al. [15] and Dodd and Pepe [2]. The dataset is from a study of the distortion product otoacoustic emissions (DPOAE) test to diagnose

Table 3 The parameters estimates in the AUC regression model $\mathrm{AUC}_{\mathbf{Z}} = \Phi(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2)$

n	m	β_0	$\hat{\beta}_0$	Bias of $\hat{\beta}_0$	sd of $\hat{\beta}_0$
30	30	0.19205	0.78011	0.58806	0.56746
50	50	0.19205	0.75420	0.56215	0.42027
100	100	0.19205	0.73639	0.54433	0.34271
n	m	β_1	$\hat{\beta}_1$	Bias of $\hat{\beta}_1$	sd of $\hat{\beta}_1$
30	30	−0.32009	−0.44119	−0.12110	0.02658
50	50	−0.32009	−0.41946	−0.09937	0.01459
100	100	−0.32009	−0.40996	−0.08987	0.01040
n	m	β_2	$\hat{\beta}_2$	Bias of $\hat{\beta}_2$	sd of $\hat{\beta}_2$
30	30	0.89625	0.93949	0.04323	0.03806
50	50	0.89625	0.89340	−0.00285	0.01471
100	100	0.89625	0.87085	−0.02540	0.00701

Table 4 Coverage probabilities of 90% and 95% confidence region for the parameters vector in the AUC regression model $\mathrm{AUC}_{\mathbf{Z}} = \Phi(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2)$

Level	n	m	HBEL1	HBEL2	NA
90%	30	30	0.891	0.887	0.792
	50	50	0.897	0.889	0.556
	100	100	0.897	0.893	0.214
95%	30	30	0.945	0.927	0.864
	50	50	0.946	0.932	0.686
	100	100	0.941	0.930	0.307

the hearing impairment. The study involved 107 hearing impaired and 103 normally hearing subjects who were examined at three frequency (f) and three intensity (L) settings of the DPOAE device. An audiometric threshold can be yielded at each setting. If the audiometric threshold is greater than 20 dB HL, the disease variable $D = 1$; otherwise $D = 0$. Each subject was tested in only one ear. The test result is the negative signal to noise ratio, $-$SNR. The covariates are selected to be X_f = frequency HZ/100, X_L = intensity dB/10, X_D = (hearing threshold $-$ 20)dB/10,

The model of interest is

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

Using the usual GLM-based estimation method, we obtain the estimate for the parameter vector in the AUC regression as follows:

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (4.22572, 0.695502, -0.82614, 0.052909).$$

It is shown that increasing the hearing threshold appears to increase the AUC odds by 100.47% for every 10 dB increase (AUC odds, 2.0047), the AUC odds decrease by 57% for every 10 dB increase in intensity (AUC odds, 0.4377), and the AUC odds increase 5.4% for every 100 HZ increase in the frequency. These results coincide with the results of [2].

Based on the proposed HBEL1 and HBEL2 methods, we can construct 95% confidence regions for the parameters vector $\boldsymbol{\beta}$ in the AUC regression as follows:

95% HBEL1 region: $\text{CR}_1(\boldsymbol{\beta}) = \{\boldsymbol{\beta} : l(\boldsymbol{\beta}) \leqslant 4511.467\}$,

95% HBEL2 region : $\text{CR}_2(\boldsymbol{\beta}) = \{\boldsymbol{\beta} : l(\boldsymbol{\beta}) \leqslant 4425.534\}$.

Using above confidence regions for $\boldsymbol{\beta}$ and the method proposed in Subsection 3.2, we can construct confidence intervals for the summary index AUC at a specified value of the covariate vector $\mathbf{Z}_D = (X_D, X_L, X_f)$. Here, we take the specified value of the covariate vector $\mathbf{Z}_D$ to be the median of the observed values for the covariate vector $\mathbf{Z}_D$, and choose $J = 200$. At the median covariates, the estimate for the covariate-specific AUC is 0.9891, the 95% HBEL1 interval for the AUC is $(0.8163, 1)$, and the 95% HBEL2 interval for the AUC is $(0.8245, 1)$. These intervals indicate that the test has moderate to high diagnostic accuracy when the covariate vector $\mathbf{Z}_D$ is set to be the median value of the covariates.

6 Discussion

Pepe and Cai [10] proposed a pseudo likelihood-based inference in the ROC regression model. However, to our knowledge, there is no likelihood-based inference in the AUC regression model. In this article, empirical likelihood-based inferences are proposed for the AUC regression model. The proposed method has sound theoretical property and can be used to construct confidence regions/intervals for the regression parameter vector and a covariate-specific AUC. The simulation results have shown that the new method outperforms the existing method. The newly proposed method combines the power of the likelihood-based approach with the pragmatic need to find a tractable and easily implemented solution in the AUC regression analysis. The method developed here is a great addition to the existing method for the AUC regression model. It also should be pointed out that the EL-based inferences can similarly be made in the ROC regression model. Future research will focus on the comparative study of the EL-based inference with the existing pseudo likelihood-based inference in the ROC regression model.

Acknowledgements The authors would like to thank the associate editor and the referees for their valuable comments which led to substantial improvements in this article. This work was supported by Southwest Jiao Tong University (Grant Nos. 12BR030 and 12ZT15), US National Science Foundation (Grant No. MPS/DMS 0603913) and US National Security Agency (Grant No. H98230-12-1-0228).

References

1 Bamber D C. The area above the ordinal dominance graph and the area below the receiver operating characteristic curve graph. J Math Psychol, 1975, 12: 387–415
2 Dodd L E, Pepe M S. Partial AUC estimation and regression. Biometrics, 2003, 59: 614–623
3 Dorfman D D, Berbaum K S, Metz C E. Receiver operating characteristic analysis: generalization to the population of readers and patients with the jackknife method. Invest Radio, 1992, 27: 723–731
4 Hall P, La Scala B. Methodology and algorithms of empirical likelihood. Int Stat Rev, 1990, 58: 109–127
5 Lin H, Zhou X H, Li G. A direct semiparametric receiver operating characteristic curve regression with unknown link and baseline functions. Stat Sinica, Preprint, doi:10.5705/ss.2010.167
6 Obuchowski N A. Multireader, multimodality receiver operating characteristic curve studies: Hypothesis testing and sample size estimation using analysis of variance approach with dependent observations. Acad Radiol, 1995, 2: S22–S29
7 Owen A. Empirical Likelihood. Noca Raton: Chapman Hall, CRC, 2001
8 Owen A. Empirical likelihood ratio confidence intervals for single functional. Biometrika, 1988, 75: 237–249
9 Owen A. Empirical likelihood ratio confidence regions. Ann Stat, 1990, 18: 90–120
10 Pepe M S, Cai T. The analysis of placement values for evaluating discriminatory measures. Biometrics, 2004, 60: 528–535
11 Pepe M S. The Statistical Evaluation of Medical Tests for Classification and Prediction. New York: Oxford University Press, 2003
12 Pepe M S. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. Biometrics, 1998, 54: 124–135
13 Qin G S, Jing B Y. Empirical likelihood for censored linear regression. Scand J Stat, 2001, 28: 661–673
14 Qin G S, Zhou X H. Empirical likelihood inference for the area under the ROC curve. Biometrics, 2006, 62: 613–622
15 Stover L, Gorga M P, Neely S T, et al. Toward optimizing the clinical utility of distortion product otoacoustic emission measurements. JASA, 1996, 100: 956–967
16 Swets J A, Pickett R M. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. New York: Academic Press, 1982
17 Thompson M L, Zucchini W. On the statistical analysis of ROC curves. Stat Med, 1989, 8: 1277–1290
18 Tosteson A N A, Begg C B. A general regression methodology for ROC curve estimation. Med Decis Making, 1988, 8: 204–215
19 Zhou X H, Obuchowski N A, McClish D K. Statistical Methods in Diagnostic Medicine. New York: John Wiley, Sons, 2002
20 Zhou X H, Qin G S, Lin H Z, et al. Inferences in censored cost regression models with empirical likelihood. Stat Sinica, 2006, 16: 1213–1232

Appendix

We need Lemmas 1 and 2 for the proof of Theorem 1.

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

Lemma 1. *Under the conditions in Theorem* 1*, we have* $\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\widehat{\mathbf{H}}_j \xrightarrow{\mathcal{L}} N(0, \boldsymbol{V})$. *where*

$$\boldsymbol{V} = \boldsymbol{V}_1 + \boldsymbol{V}_2,$$

$$\boldsymbol{V}_1 = \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}(\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D)^{\otimes 2}\mathrm{Var}_{\mathbf{Z}^D}^{D}[F_{\mathbf{Z}^{\bar{D}}}^{\bar{D}}(Y^D)],$$

$$\boldsymbol{V}_2 = \rho\bigg(\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D\bigg)^{\otimes 2}\mathrm{Var}_{\mathbf{Z}^{\bar{D}}}^{\bar{D}}\bigg[\int I(Y^{\bar{D}} \leqslant t)dF_{\mathbf{Z}^{\mathbf{D}}}^{D}(t)\bigg].$$

Proof. From

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\widehat{\mathbf{H}}_j = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}[1-\widehat{U}_j^D - g(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)]\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D,$$

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbf{H}_j = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}[1-U_j^D - g(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)]\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D,$$

we obtain the following decomposition:

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\widehat{\mathbf{H}}_j=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbf{H}_j+\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(U_j^D-\widehat{U}_j^D)\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D\equiv I_1+I_2.$$

From

$$\begin{aligned}\mathrm{Var}(I_1)&=\frac{1}{n}\mathrm{Var}\bigg(\sum_{j=1}^{n}[F_{\mathbf{Z}_{\bar{D}}}^{\bar{D}}(Y_j^D)-g(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)]\boldsymbol{\omega}(\boldsymbol{\beta}_0^T\mathbf{Z}_j^D)\mathbf{Z}_j^D\bigg)\\&=\frac{1}{n}\sum_{j=1}^{n}(\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D)^{\otimes 2}\mathrm{Var}_{\mathbf{Z}^D}^D[F_{\mathbf{Z}_{\bar{D}}}^{\bar{D}}(Y_j^D)]\\&\longrightarrow\boldsymbol{V}_1\end{aligned}$$

(If $\mathbf{Z}_j^D$'s are i.i.d. random variables, then $\boldsymbol{V}_1=E((\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}^D)\mathbf{Z}^D)^{\otimes 2}\mathrm{Var}_{\mathbf{Z}^D}^D[F_{\mathbf{Z}^{\bar{D}}}^{\bar{D}}(Y^D)]))$, and Central Limit Theorem, it follows that $I_1\xrightarrow{\mathcal{L}}N(0,\boldsymbol{V}_1)$.

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

$$\begin{aligned}I_2&=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\widehat{F}_{\mathbf{Z}^{\bar{\mathbf{D}}}}^{\bar{D}}(Y_j^D)-F_{\mathbf{Z}^{\bar{\mathbf{D}}}}^{\bar{D}}(Y_j^D))\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D\\&=\bigg[\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D\bigg]\int\sqrt{n}(\widehat{F}_{\mathbf{Z}^D}^{\bar{D}}(t)-F_{\mathbf{Z}^D}^{\bar{D}}(t))dF_{\mathbf{Z}^D}^D(t)+o_p(1)\\&=\sqrt{\frac{n}{m}}\bigg[\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D\bigg]\times\sqrt{m}\bigg(\frac{1}{m}\sum_{i=1}^{m}\int I(Y_i^{\bar{D}}\leqslant t)dF_{\mathbf{Z}^D}^D(t)-\int F_{\mathbf{Z}^{\bar{\mathbf{D}}}}^{\bar{D}}(t)dF_{\mathbf{Z}^D}^D(t)\bigg)+o_p(1)\\&\xrightarrow{\mathcal{L}}N(0,\boldsymbol{V}_2).\end{aligned}$$

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, $\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\widehat{\mathbf{H}}_j=I_1+I_2\xrightarrow{\mathcal{L}}N(0,\boldsymbol{V})$, with $\boldsymbol{V}=\boldsymbol{V}_1+\boldsymbol{V}_2$.

Lemma 2. *Under the conditions in Theorem* 1*, we have that*

(i) $\max_j\|\widehat{\mathbf{H}}_j\|=o_p(n^{1/2})$;

(ii) $\boldsymbol{V}_{0n}=\frac{1}{n}\sum_{j=1}^{n}\widehat{\mathbf{H}}_j^{\otimes 2}\xrightarrow{p}\boldsymbol{V}_0=\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\mathbf{H}_j^{\otimes 2}$.

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

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

it follows that

$$\begin{aligned}\|\widehat{\boldsymbol{H}}_j-\boldsymbol{H}_j\|&=\|(U_j^D-\widehat{U}_j^D)\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D\|\\&=|\widehat{F}_{\mathbf{Z}^{\bar{D}}}^{\bar{D}}(Y_j^D)-F_{\mathbf{Z}^{\bar{D}}}^{\bar{D}}(Y_j^D)|\|\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D\|\\&\leqslant\sup_y|\widehat{F}_{\mathbf{Z}^{\bar{D}}}^{\bar{D}}(y)-F_{\mathbf{Z}^{\bar{D}}}^{\bar{D}}(y)|\|\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D\|=o_p(1),\end{aligned}$$

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

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

(ii) Denote $\widetilde{\boldsymbol{V}}_{0n}=\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{H}_j^{\otimes 2}$, where $\boldsymbol{H}_j=(1-U_j^D-g(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D))\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D$. For any fixed $\boldsymbol{a}\in\mathbb{R}^p$,

$$\boldsymbol{a}^{\mathrm{T}}(\boldsymbol{V}_{0n}-\widetilde{\boldsymbol{V}}_{0n})\boldsymbol{a}=\frac{1}{n}\sum_j(\boldsymbol{a}^{\mathrm{T}}(\widehat{\boldsymbol{H}}_j-\boldsymbol{H}_j))^2+\frac{2}{n}\sum_j(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{H}_j)(\boldsymbol{a}^{\mathrm{T}}(\widehat{\boldsymbol{H}}_j-\boldsymbol{H}_j))$$

$$\leqslant \frac{1}{\sqrt{n}}\sum_j |\boldsymbol{a}^{\mathrm{T}}(\widehat{\boldsymbol{H}}_j-\boldsymbol{H}_j)|\cdot\Big(\frac{1}{\sqrt{n}}\max_j|\boldsymbol{a}^{\mathrm{T}}(\widehat{\boldsymbol{H}}_j-\boldsymbol{H}_j)|+\frac{2}{\sqrt{n}}\max_j|\boldsymbol{a}^{\mathrm{T}}\boldsymbol{H}_j|\Big)$$
$$= J_0\cdot(J_1+2J_2).$$

Since

$$J_0=\frac{1}{\sqrt{n}}\sum_j|\boldsymbol{a}^{\mathrm{T}}(\widehat{\boldsymbol{H}}_j-\boldsymbol{H}_j)|=\frac{1}{\sqrt{n}}\sum_{j=1}^n|\widehat{F}^{\bar{D}}_{\mathbf{Z}_{\bar{\mathbf{D}}}}(Y_j^D)-F^{\bar{D}}_{\mathbf{Z}_{\bar{\mathbf{D}}}}(Y_j^D)||\boldsymbol{a}^{\mathrm{T}}\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D|$$
$$\leqslant \sup_y|\sqrt{n}(\widehat{F}^{\bar{D}}_{\mathbf{Z}_{\bar{\mathbf{D}}}}(y)-F^{\bar{D}}_{\mathbf{Z}_{\bar{\mathbf{D}}}}(y))|\cdot\frac{1}{n}\sum_{j=1}^n|\boldsymbol{a}^T\omega(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_j^D)\mathbf{Z}_j^D|=O_p(1),$$
$$J_1=\frac{1}{\sqrt{n}}\max_j|\boldsymbol{a}^{\mathrm{T}}(\widehat{\boldsymbol{H}}_j-\boldsymbol{H}_j)|=O\Big(n^{-1/2}\Big(\max_j\|\widehat{\boldsymbol{H}}_j\|+\max_j\|\boldsymbol{H}_j\|\Big)\Big)=o_p(1),$$
$$J_2=\frac{1}{\sqrt{n}}\max_j|\boldsymbol{a}^{\mathrm{T}}\boldsymbol{H}_j|=O\Big(n^{-1/2}\max_j\|\boldsymbol{H}_j\|\Big)=o_p(1),$$

then $\boldsymbol{V}_{0n}=\tilde{\boldsymbol{V}}_{0n}+o_p(1)$. Lemma 2 is thus proved. □

Proof of Theorem 1. Using Lemmas 1 and Lemma 2 (ii) and the similar argument used in [9], we can prove that

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

Then, applying Taylor's expansion, we get that

$$l(\boldsymbol{\beta}_0)=2\sum_{j=1}^n\log(1+\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j)=2\sum_{j=1}^n\Big(\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j-\frac{1}{2}(\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j)^2\Big)+r_{1n},$$

with

$$|r_{1n}|\leqslant C\sum_{j=1}^n|\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j|^3\leqslant C\|\boldsymbol{\nu}\|^3\max_j\|\widehat{\mathbf{H}}_j\|\sum_j\|\widehat{\mathbf{H}}_j\|^2=o_p(1).$$

By (2.4), we have

$$\sum_{j=1}^n\frac{\widehat{\mathbf{H}}_j}{1+\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j}=\sum_{j=1}^n\widehat{\mathbf{H}}_j\Big[1-\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j+\frac{(\boldsymbol{\nu}\widehat{\mathbf{H}}_j)^2}{1+\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j}\Big]$$
$$=\sum_{j=1}^n\widehat{\mathbf{H}}_j-\Big(\sum_{j=1}^n\widehat{\mathbf{H}}_j^{\otimes2}\Big)\boldsymbol{\nu}+\sum_{j=1}^n\frac{\widehat{\mathbf{H}}_j(\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j)^2}{1+\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j}=0. \tag{A.2}$$

(A.1), (A.2) and Lemma 2 together imply that

$$\boldsymbol{\nu}=\Big(\sum_{j=1}^n\widehat{\mathbf{H}}_j^{\otimes2}\Big)^{-1}\sum_{j=1}^n\widehat{\mathbf{H}}_j+o_p(n^{-1/2}).$$

Again by (2.4), we get that

$$0=\sum_{j=1}^n\frac{\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j}{1+\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j}=\sum_{j=1}^n(\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j)\Big[1-\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j+\frac{(\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j)^2}{1+\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j}\Big]$$
$$=\sum_{j=1}^n(\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j)-\sum_{j=1}^n(\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j)^2+\sum_{j=1}^n\frac{(\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j)^3}{1+\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j}. \tag{A.3}$$

From (A.1) and Lemma 2, we can get

$$\frac{1}{n}\sum_{j=1}^n\frac{(\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j)^3}{1+\boldsymbol{\nu}^{\mathrm{T}}\widehat{\mathbf{H}}_j}=o_p(1),$$

then we get

$$\sum_{j=1}^{n} \boldsymbol{\nu}\widehat{\mathbf{H}}_j = \sum_{j=1}^{n} (\boldsymbol{\nu}\widehat{\mathbf{H}}_j)^2 + o_p(1). \tag{A.4}$$

From (A.1)–(A.4), and Lemma 1, it follows that

$$\begin{aligned} l_1(\beta_0) &= \sum_{j=1}^{n} \boldsymbol{\nu}^{\mathrm{T}} \widehat{\mathbf{H}}_j^{\otimes 2} \boldsymbol{\nu} + o_p(1) \\ &= \left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \widehat{\mathbf{H}}_j\right)^{\mathrm{T}} \left(\frac{1}{n} \sum_{j=1}^{n} \widehat{\mathbf{H}}_j^{\otimes 2}\right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \widehat{\mathbf{H}}_j\right) + o_p(1) \\ &= \left(\frac{1}{\sqrt{n}} \boldsymbol{V}^{-1/2} \sum_{j=1}^{n} \widehat{\mathbf{H}}_j\right)^{\mathrm{T}} (\boldsymbol{V}^{1/2} \boldsymbol{V}_0^{-1} \boldsymbol{V}^{1/2}) \left(\frac{1}{\sqrt{n}} \boldsymbol{V}^{-1/2} \sum_{j=1}^{n} \widehat{\mathbf{H}}_j\right) + o_p(1). \end{aligned}$$

Note that $\boldsymbol{V}^{1/2}\boldsymbol{V}_0^{-1}\boldsymbol{V}^{1/2}$ has the same eigenvalue as $\boldsymbol{V}_0^{-1}\boldsymbol{V}$. We denote $\boldsymbol{V}_0^{-1}\boldsymbol{V}$ as $\mathbf{G}$. Then the conclusion of Theorem 1 follows from Lemmas 1, 2 (ii) and 3 in [13]. □