Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Semi-empirical likelihood inference for the ROC curve with missing data

Xiaoxia Liu [a,*], Yichuan Zhao [b]

[a] Department of Anesthesiology, Perioperative and Pain Medicine, Brigham and Women's Hospital, Boston, MA 02115, United States
[b] Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, United States

ARTICLE INFO

Article history:
Received 28 November 2010
Received in revised form
22 May 2012
Accepted 13 June 2012
Available online 26 June 2012

Keywords:
Confidence interval
Empirical likelihood
Estimating equation
Hot deck imputation
ROC curve

ABSTRACT

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

1. Introduction

In medical research, the receiver operating characteristic (ROC) curve analysis has been extensively used in the evaluation of diagnostic tests. Generally speaking, ROC curve is the entire set of possible true and false positive fractions attained by dichotomizing a continuous test result T with different thresholds (Pepe, 2003). That is, the ROC curve is $\mathrm{ROC}(\cdot)=\{(\mathrm{FPF}(c),\mathrm{TPF}(c)),c\in(-\infty,\infty)\}$. Alternatively, for a continuous-scale diagnostic test, let X and Y be the test results from diseased subjects and non-diseased subjects respectively. At a given cutoff point c, the sensitivity and specificity are defined as $\mathrm{Se}=\Pr(X\geq \mathrm{c})$ and $\mathrm{Sp}=\Pr(Y<c)$. If $F(\cdot)$ and $G(\cdot)$ are the corresponding distribution functions of X and Y, the sensitivity and specificity can then be written as $\mathrm{Se}=1-F(c)$ and $\mathrm{Sp}=G(c)$. Then ROC curve is actually a plot of $1-F(c)$ versus $1-G(c)$, for $-\infty<c<\infty$. At a fixed level $q=(1-$ specificity), the ROC curve can be expressed as $\Delta_q=1-F\{G^{-1}(1-q)\}$, for $0<q<1$, where G^{-1} is the inverse function of G, i.e., $G^{-1}(q)=\inf\{c:G(c)\geq q\}$.

Varieties of approaches regarding estimation of ROC curve have been developed, both parametric and non-parametric. Over the years, confidence interval for a continuous-scale ROC curve has received much attention because it is more useful than point estimates and more helpful for researchers to make accurate diagnostic decisions (Su et al., 2009). To avoid the deficiencies of normal approximation for small or moderate sample sizes, the empirical likelihood (EL) based method is extensively used to estimate Δ_q. EL is a non-parametric way of inference based on a data-driven likelihood ratio function. The first use of empirical likelihood ratio function to get confidence intervals appears to be Thomas and Grunkemeier

[*] Corresponding author.
E-mail addresses: xiaoxia.liu@gmail.com (X. Liu), yichuan@gsu.edu (Y. Zhao).

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2012.06.011

(1975). They showed that empirical likelihood ratio confidence intervals for a survival probability based on the χ_1^2 distribution have asymptotically correct coverage levels. Later on, the empirical likelihood methods for constructing confidence regions for the mean parameter of the population were developed systematically by Owen (1988) and Owen (1990). Comparing to normal approximation method and bootstrap method, EL method can improve the confidence region, and increase accuracy of the coverage (Hall and La Scala, 1990). Recently, much attention has been paid to smoothing strategies to overcome the discontinuity issue of the ROC curves. Chen and Hall (1993) first of all introduced smoothed EL confidence intervals for quantiles on one population. Zou et al. (1997), Lloyd (1998) and Ren et al. (2004) proposed various smoothed estimators for ROC curves among others. Liang and Zhou (2008) then examined semiparametric empirical likelihood confidence intervals for ROC curves with right censoring and established the asymptotic result. More recently, Yang and Zhao (2012) developed smoothed EL confidence intervals for the ROC curve with right censoring. The principle of our smoothed empirical likelihood is similar to that of Chen and Hall (1993) and Claeskens et al. (2003) in the spirit.

Current available methods in analyzing ROC curves are limited to complete data regardless of parametric or non-parametric settings. The EL method needs modifications when dealing data with missing values. The procedure in our context is different from that for usual situations. In this paper, we extend the previous studies and concern the situation that when one model is parametric while the other one is non-parametric, both with missing data in them. As a matter of fact, this is a very common case in medical research or related fields. For example, when comparing a new treatment with control treatment, we tend to have more if not enough information about the well developed treatment (i.e., control treatment), while the new treatment is less known. This leads to a semi-parametric two-sample model, which can reflect the difference of two samples of missing data. Let X and Y be the responses of two samples, for example, the diseased and non-diseased subjects, and we assume $F(\cdot)$ and $G(\cdot)$ are the distribution functions of X and Y respectively. Furthermore, we have the assumption that the population distribution function F is non-parametric while G is parametric, and both X and Y with missing data in them. In this paper, we are interested in constructing confidence intervals for the ROC curve, or Δ_q with missing data under this specific context by using empirical likelihood ratio methods. We are interested in establishing asymptotic distribution of the resulting statistics and derive empirical likelihood-based confidence intervals for the Δ_q. We also prove that the log likelihood ratio is asymptotically scaled chi-squared distribution.

The organization of this paper is as follows. In Section 2 we introduce the hot deck imputation method first. Then the smoothed empirical likelihood for the ROC curve is proposed. Also, the semi-empirical likelihood based confidence interval is constructed and the asymptotic results are established. In Section 3, simulation studies are conducted to evaluate the finite sample performance of the proposed method. In Section 4 we give a summary and discussion. All proofs are included in the Appendix.

2. Inference procedure

2.1. Missing data imputation

Throughout the paper, we adopt similar notations as Qin and Zhang (2009). Consider the following random samples associated with two independent populations (x,δ_x) and (y,δ_y):

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

where missing indicators

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

We assume missing completely at random (MCAR). Put $P(\delta_x=1|x)=P_1$ and $P(\delta_y=1|y)=P_2$. Like Qin and Zhang (2009), denote $r_x=\sum_{i=1}^m \delta_{xi}$, $r_y=\sum_{j=1}^n \delta_{yj}$. The respondents with respect to x and y can be written as s_{rx} and s_{ry}, respectively. While the non-respondents are denoted as s_{mx} and s_{my} corresponding to x and y, respectively, where $m_x=m-r_x$ and $m_y=n-r_y$. Let x_i^* and y_j^* denote the imputed values for the missing data with respect to x and y, respectively.

Imputation is a commonly used technique to handle missing data which does not strive to determine the best predictions of individual missing values. Instead, imputation is the substitution of plausible values for missing data so that inference about the parameters of interest can be made using retained information from the incomplete observations (Andridge and Little, 2010). As a result, there is a possible gain in efficiency comparing to a complete data analysis, and a reduction in non-response bias (Little and Rubin, 2002).

For the sample X, which comes from a non-parametric population, we impute the missing values. We adopt here random hot deck imputation method which is less sensitive to model misspecification and only imputes plausible values since they are from observed responses. Also, the hot deck estimator is unbiased under MCAR assumption (Little and Rubin, 2002). For the sample Y, which comes from a parametric population, we first get maximum likelihood estimator (MLE) of population parameter θ, then select random samples from the population with this estimated parameter. Let $\hat{\theta}$ denote the MLE of θ from the sample $\{y_j, j \in s_{ry}\}$. Then we choose a random sample with size m_y from the population $G_{\hat{\theta}}(\cdot)$