Journal of Statistical Planning and Inference 140 (2010) 1162–1174

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Non-parametric estimation for time-dependent AUC

Chin-Tsang Chiang *, Hung Hung

Department of Mathematics, National Taiwan University, Taipei 10617, Taiwan, ROC

ARTICLE INFO

Article history:
Received 6 January 2009
Received in revised form
22 October 2009
Accepted 28 October 2009
Available online 3 November 2009

Keywords:
AUC
Bivariate estimation
Bootstrap
Gaussian process
Kaplan–Meier estimator
Non-parametric estimator
ROC
Smoothing parameter
Survival data

ABSTRACT

The area under the receiver operating characteristic (ROC) curve (AUC) is one of the commonly used measure to evaluate or compare the predictive ability of markers to the disease status. Motivated by an angiographic coronary artery disease (CAD) study, our objective is mainly to evaluate and compare the performance of several baseline plasma levels in the prediction of CAD-related vital status over time. Based on censored survival data, the non-parametric estimators are proposed for the time-dependent AUC. The limiting Gaussian processes of the estimators and the estimated asymptotic variance–covariance functions enable us to further construct confidence bands and develop testing procedures. Applications and finite sample properties of the proposed estimation methods and inference procedures are demonstrated through the CAD-related death data from the British Columbia Vital Statistics Agency and Monte Carlo simulations.

1. Introduction

Let T be the time to the death or the diagnosis of a specific disease with the corresponding time-varying vital or disease status I ($T \le t$), and Y denote the continuous marker measured at or before the outset of study. Heagerty et al. (2000) defined the time-dependent true positive rate $TPR(y,t) = P(Y > y | T \le t)$ and false positive rate $FPR(y,t) = P(Y > y | T > t)$ for any varying threshold value y. It is easily to show that $TPR(y,t)$ and $FPR(y,t)$ can be expressed as

$$TPR(y,t) = \frac{S_Y(y) - S(t,y)}{1 - S_T(t)} \quad \text{and} \quad FPR(y,t) = \frac{S(t,y)}{S_T(t)}, \tag{1}$$

where $S(t,y) = P(T > t, Y > y)$ with $S_Y(y) = S(0,y)$ and $S_T(t) = S(t,-\infty)$ being the survival functions of Y and T. The time-dependent ROC curve, denoted by ROC_t, displays the spectrum of values for $TPR(y,t)$ against $FPR(y,t)$ over varying threshold values y. By integrating the area under ROC_t, the time-dependent AUC θ_t is directly obtained as

$$\theta_t = \frac{-\int (S_Y(y) - S(t,y))\, d_y S(t,y)}{S_T(t)(1 - S_T(t))} = \frac{-\int (S_Y(y) - S(t,y))\, dS_Y(y) - 0.5(1 - S_T(t))^2}{S_T(t)(1 - S_T(t))}, \tag{2}$$

where $d_y S(t,y)$ is the Lebesgue–Stieltjes integration over y for fixed t. The above formulation is found to be more successful in seeking estimators and developing inference procedures than the probability expression $P(Y_i > Y_j | T_i \le t, T_j > t)$, $i \ne j$, of θ_t, which can be derived based on a random sample $\{(T_i, Y_i)\}_{i=1}^n$ from a homogeneous distribution. Different from the time-invariant AUC, the cases (diseased or dead individuals) and controls (non-diseased or alive ones) in the classification

* Corresponding author.
E-mail address: chiangct@ntu.edu.tw (C.-T. Chiang).

0378-3758/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2009.10.012

probability θ_t are defined over time. The capacity of a marker is monitored at any time point of interest to accurately classify individuals as cases or controls. The greatest challenge for this issue is that the vital statuses of some individuals might not be available due to censoring.

In an angiography cohort study, Lee et al. (2006) detected that the elevated plasma levels of C-reactive protein (CRP), serum amyloid A protein (SAA), Interleukin(IL)-6, and total homocysteine (tHcy) are linked to CAD-related event. It is worthwhile to further evaluate and compare the predictive abilities of these four plasma biomarkers on the CAD-related vital status over time instead of at the end of study. From the British Columbia Vital Statistics data, the vital statuses of some patients might not be available due to censoring. Thus, the censored survival data $\{(X_i, \delta_i, Y_i)\}_{i=1}^n$ are considered in this article, where $X_i = \min\{T_i, C_i\}$ is the follow-up time with C_i being the censoring time and $\delta_i = I(X_i = T_i)$ denotes the censoring status, $i = 1, \ldots, n$. Although an estimation method for θ_t can be obtained via computing an appropriate area under the estimate of ROC_t, there is still no rigorous theoretical basis for this method. To estimate θ_t directly from the data, Chambless and Diao (2006) proposed a non-parametric recursive estimation method, which only provides estimates on the failure times of observed data, and a model based estimation one, which has a similar formulation as (2), in terms of a conditional survival estimator. In contrast to the recursive estimator, our proposed non-parametric estimators are well defined at any time point within the study period. The proposed methods are mainly based on (2) and the non-parametric bivariate estimators of $S(t, y)$ derived under a very suitable assumption of conditionally independent censoring (A1: Conditioning on Y, C and T are independent). Our methods provide easily computed estimates of θ_t and can be successfully applied to left truncation and right censoring data. With some mild conditions, the asymptotic Gaussian processes of the estimators are established. Moreover, the uniform asymptotic representations of the estimators enable us to propose uniformly consistent estimators for the asymptotic variance–covariance functions, construct confidence bands for θ_t, and develop the testing procedures for the hypothesis of equality of several time-dependent AUC curves over time.

This paper is organized as follows. In Section 2, the non-parametric estimation methods for θ_t and the asymptotic variance–covariance functions of the estimators are proposed. The corresponding statistical inferences are also presented in this section. Section 3 conducts a class of simulation studies to examine the finite sample properties of the estimators and the performance of the proposed inference procedures. An application of our methods to an angiography CAD study is provided in Section 4. We conclude with a brief discussion in Section 5 and place the proofs of the main results in the Appendix.

2. Estimation methods and statistical inferences

In this section, the non-parametric estimation methods are proposed for θ_t. The asymptotic Gaussian process results and the estimated asymptotic variance–covariance functions enable us to construct confidence bands for θ_t and compare the utilities of several baseline plasma biomarkers.

2.1. Non-parametric estimators for θ_t

As in an angiography CAD study, the censoring times of patients might be affected by the baseline plasma biomarkers. The assumption of random censoring might be unreasonable and can be relaxed towards a more general assumption of conditionally independent censorship (A1). Let $S_X(t|y) = P(X > t|Y = y)$ and $S_{X\delta}(t|y) = P(X > t, \delta = 1|Y = y)$. These conditional functions are suggested to be estimated by the smoothing estimators $\tilde{S}_{X,\lambda}(t|y) = n^{-1}\sum_{i=1}^n I(X_i \geq t)K_1(\lambda^{-1}\{\hat{S}_Y(Y_i) - \hat{S}_Y(y)\})$ and $\tilde{S}_{X\delta,\lambda}(t|y) = n^{-1}\sum_{i=1}^n \delta_i I(X_i \geq t)K_1(\lambda^{-1}\{\hat{S}_Y(Y_i) - \hat{S}_Y(y)\})$, where $\hat{S}_Y(y)$ is an empirical estimator of $S_Y(y)$, $K_1(u) = 0.5I(|u| < 1)$, and λ is a non-negative smoothing parameter. Thus, the bivariate function $S(t, y)$ is suggested to be estimated by the nearest neighbor estimator $\tilde{S}_\lambda(t, y) = -\int I(u > y)\tilde{S}_{T,\lambda}(t|u)d\hat{S}_Y(u)$ of Akritas (1994) with

$$\tilde{S}_{T,\lambda}(t|y) = \mathcal{P}_0^t\left\{1 + \frac{d_u\tilde{S}_{X\delta,\lambda}(u|y)}{\tilde{S}_{X,\lambda}(u|y)}\right\} \tag{3}$$

being an estimator of $S_T(t|y) = P(T > t|Y = y)$ and $\mathcal{P}_0^t$ denoting the infinite product integral. An alternative smoothing estimator for $S_T(t|y)$ can also be found in Dabrowska (1987a) with the weights $K_2(\lambda^{-1}\{Y_i - y\})$'s being adopted in (3), where $K_2(\cdot)$ is some probability density function. The kernel weight function used in this article is more suitable because it requires fewer smoothness conditions (cf. Akritas, 1994) and the choice of smoothing parameter is not affected by the measurement scale of Y. By substituting $\tilde{S}_\lambda(t, y)$, $\tilde{S}_{T,\lambda}(t) = \tilde{S}_\lambda(t, -\infty)$, and $\hat{S}_Y(y)$ separately for $S(t, y)$, $S_T(t)$, and $S_Y(y)$ in the first equality of (2), the estimator $\tilde{\theta}_t$ for θ_t is obtained. Moreover, θ_t can be estimated in different ways using the second equality of (2) and this alternative estimator is denoted by $\tilde{\theta}_{2t}$. Adirect calculation shows that $\tilde{\theta}_t$ has the following expression:

$$\tilde{\theta}_t = \frac{n^{-2}\sum_{i\neq j}(1 - \tilde{S}_{T,\lambda}(t|Y_i))\tilde{S}_{T,\lambda}(t|Y_j)I(Y_i > Y_j)}{\tilde{S}_{T,\lambda}(t)(1 - \tilde{S}_{T,\lambda}(t))}. \tag{4}$$

The notations below are used to help describe the asymptotic properties of $\tilde{\theta}_t$: $D = (X, \delta, Y)$, $G_t = P(Y_i > Y_j, T_i > t)$, $S_{XY}(t, y) = P(X > t, Y > y)$, $N(t) = I(X \leq t)\delta$, $M(t|y) = N(t) + \int_0^t I(X \geq u)d_u(\ln S_T(u|y))$, $\xi_t(Y, X, \delta) = -S_T(t|Y)\int_0^t d_u M(u|Y)/S_X(u|Y)$, and $\eta_t = \{G_t(1 - 2S_T(t)) + 0.5S_T^2(t)\}\{S_T(t)(1 - S_T(t))\}^{-1}$. Moreover, let $\Omega = \{(t, y) : S_{XY}(t, y) > 0\}$, $(\tau_1, M_y) = \text{argmin}_{(t,y)\in\Omega}S_{XY}(t, y)$, and $\tau_0 = \inf\{u : S_T(u) < 1, u \leq \tau_1\}$. It can be derived that $n^{1/2}(\tilde{\theta}_{2t} - \tilde{\theta}_t) = 0.5n^{-1/2}\int\tilde{S}_{T,\lambda}^2(t|y)d\hat{S}_Y(y)\{\tilde{S}_{T,\lambda}(t)(1 - \tilde{S}_{T,\lambda}(t))\}^{-1} \leq 0$ and