



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

석사학위논문

Time-dependent ROC 곡선을
이용한 연속형 바이오마커의
최적의 절단점을 찾는 방법
비교: 모의실험 연구

2022년 2월 22일

전 북 대 학 교 대 학 원

통계학과

조 진 실

Time-dependent ROC 곡선을
이용한 연속형 바이오마커의
최적의 절단점을 찾는 방법
비교: 모의실험 연구

A Comparison of Determining Optimal Cutpoints
in Continuous Biomarkers Utilizing a
Time-dependent ROC Curve: A simulation study

2022년 2월 22일

전 북 대 학 교 대 학 원

통계학과

조 진 실

Time-dependent ROC 곡선을
이용한 연속형 바이오마커의
최적의 절단점을 찾는 방법
비교: 모의실험 연구

지도교수 최 혜 미

이 논문을 의학 석사 학위논문으로 제출함.

2021년 10월 20일

전 북 대 학 교 대 학 원

통계학과

조 진 실

조진실의 석사학위논문을 인준함.

위 원 장	전북대학교 조교수	양성준	(인)
부위원장	전북대학교 조교수	최규빈	(인)
위 원	전북대학교 교 수	최혜미	(인)

2022년 1월 4일

전 북 대 학 교 대 학 원

차 례

제 1 장	서 론	1
제 2 장	Time-dependent ROC 곡선 기반 방법	4
제 1 절	ROC 곡선을 이용한 절단점 추정	4
제 2 절	Time-dependent ROC 곡선을 이용한 절단점 추정	5
2.1	최근접 이웃(Nearest Neighbor: NN)방법	7
2.2	중도절단 가중치의 역확률 (Inverse Probability of Censoring Weighting: IPCW) 방법	9
2.3	비례 위험 (Proportional Hazards: PH) 방법	10
제 3 장	모의실험	12
제 1 절	자료의 생성	12
제 2 절	모의실험 설계	13
제 3 절	모의실험 결과	14
제 4 장	결론	17
부록 A	모의실험 결과 표	24

그림 차례

3.1	Density plot for Weibull Distribution of T	13
3.2	Survival times in the two groups meet proportional hazards assumption(Case 1) and present moderate departures from proportional hazards assumption(Case 2) when censoring rate 40%	15

표 차례

3.1 The distribution of survival time T given $Z = z$ 12

A.1 Survival times in the two groups meet proportional hazards
assumption(Case 1) and present moderate departures from
proportional hazards assumption(Case 2) when censoring rate
40% 25

ABSTRACT

A Comparison of Determining Optimal Cutpoints in Continuous Biomarkers Utilizing a Time-dependent ROC Curve: A simulation study

JINSIL JO

DEPARTMENT OF STATISTICS

THE GRADUATE SCHOOL

JEONBUK NATIONAL UNIVERSITY

In medical research, dichotomizing marker is widely used because it is easy to interpret and convenient. The most used method ROC curve is used to estimate the optimal cutpoint using the Youden index. However, since actual disease states and markers change over time, there is a risk of information loss. As an alternative to this, the optimal cutpoint of the marker can be estimated using the time-dependent ROC curve proposed by Heagerty et al. [2000], but, there is limitation using only the time-dependent ROC curve at a specific time point. Choi [2020] proposed the method of estimating the optimal cutpoint of a marker as a value that maximizes the sum of

the Youden index obtained from all uncensored survival times of the time-dependent ROC curve. The method for estimating the optimal cutpoint of a marker is proposed as a value at which the weighted sum of Youden index at each time point of the curve is maximum. Through simulation, the performance of the method proposed by Choi [2020] and the method proposed in this paper were compared through RMSE(Root Mean Square Error) for the censoring rate and size of various survival data, and the RMSE of the method proposed in this paper was smaller.

Keywords : dichotomization, prediction accuracy, time-dependent ROC, Youden index

제 1 장 서론

의학 연구에서 환자는 종종 특정 예후 인자와 같은 마커(marker)의 값에 따라 저위험군과 고위험군으로 분류된다. 예를 들어, 혈청 크레아티닌 수치는 신장 이식 후 거부 반응에 대한 중요한 마커일 수 있다. 환자가 이식 후 6개월 동안 급성 거부반응이 없고 혈청 크레아티닌 수치가 낮으면 투여된 스테로이드를 중단하는 경우가 많은데, 2.0mg/dL은 일반적으로 이 임상적 결정을 내리기 위한 혈청 크레아티닌의 절단점으로 사용된다. 이 경우 혈청 크레아티닌 수치가 2.0mg/dL 이하인 환자를 저위험군으로, 2.0mg/dL 이상인 환자를 고위험군으로 분류할 수 있다(Kim et al. [2019]).

마커의 이분화는 정보 손실, 검정력 감소, 및 단조롭지 않은 관계를 나타내는 문제점이 있다(Royston et al. [2006]). 그럼에도 불구하고, 마커는 임상 연구에서 널리 퍼져 있으며, 실제 임상 연구에서 해석이 쉽고 편리하기 때문에 일반적으로 사용하고 있다(Choi [2020]).

결과가 이진일 때, 연속형 변수를 이분화하여 최적의 절단점(cutpoint)을 찾는 방법은 여러가지가 있는데, 그 중 하나는 최대 카이제곱방법(maximal chi-squared method)이다. Sima and Gönen [2013]는 우중도절단인 경우를 고려한 3가지 방법을 제시하였다. 첫번째는 최대 일치도(maximal concordance)이다. 일치 확률을 최대화 하여 절단점을 선택하는 방법으로, ROC 곡선을 사용하여

Youden index를 최대화 하는 것과 동일한 방법을 보여주었다. 두번째는 두가지 일치 기반 메트릭(concordance based metric)인 C-index와 concordance probability estimate를 보여주었고, 마지막으로 세 개의 카이제곱기반 메트릭(Wald, 로그 순위 검정, 부분 우도 검정)의 방법을 제시하였다. 한편, 연속형 마커를 이분화 하여 최적의 절단점을 찾는 또 다른 방법 중 가장 많이 사용하는 방법은 ROC 곡선이다.

ROC 곡선은 양성 판정을 절단점을 $-\infty$ 에서 ∞ 로 변화시킬 때 민감도와 1-특이도의 변화를 그래프로 파악하는 통계기법이다. 이러한 면에서 ROC곡선은 진단검사 의학 분야에서 검사의 정확성을 파악하는 데 가장 많이 사용되고 있다(Begg [1991]). 그러나 ROC 곡선은 시간이 고정되어 있다고 가정하여 개인의 질병 상태를 평가하지만, 실제로 질병 상태와 마커값 모두 시간에 따라 변화하기 때문에 정보 손실의 위험이있다(Heagerty et al. [2000],Heagerty and Zheng [2005],Kamarudin et al. [2017]).

이에 대한 대안으로 시간에 따라 변화하는 질병 상태 및 마커값을 고려한 time-dependent ROC 곡선 방법이 소개 되었다(Heagerty et al. [2000]). 그러나 time-dependent ROC 곡선은 특정 시점에서만 정의할 수 있다는 한계로 인해 기존에 사용되는 방법인 로그 순위 검정 방법이 더 일반적으로 사용되고 있다. 이에 따라 Choi [2020] 는 time-dependent ROC 곡선의 중도절단되지 않은 모든 생존시간에서 구한 Youden 계수들의 합이 최대가 되는 절단점을 마커의 최적 절단점으로 추정하는 방법을 제안하여 주어진 정보를 충분히 활용하고자 하였다.

본 연구에서는 Choi [2020]방법을 수정한 time-dependent ROC 곡선의 각 시점별 Youden 계수들의 가중합이 최대가 되는 값으로 최적 절단점을 추정하는 방법을 제안하였다. 이 때 가중치는 생존시간의 확률밀도함수의 비모수적 추정치인 중도절단을 고려한 커널함수 추정치를 사용하였다. 모의실험을 통하여 다양한 생존 자료의 중도절단 비율 및 생존 자료의 크기의 경우에 대하여 본 논문에서 제안한 방법과 Choi [2020]의 방법을 모의실험을 통해 비교해보고자 한다.

본 논문의 구성은 2장에서 time-dependent ROC 곡선의 기반 방법들을 소개하고, 3장에서 모의실험을 진행하였으며 그에 대한 결론을 4장에 기술하였다.

제 2 장 Time-dependent ROC 곡선 기반 방법

제 1 절 ROC 곡선을 이용한 절단점 추정

의학연구에서 검사의 진단 능력을 평가하기 위해 민감도(sensitivity), 특이도(specificity), 위양성률(false positive rate), 위음성률(false negative rate), 양성예측도(true-positive), 음성예측도(true-negative), 정확도(accuracy)등을 보고, 검사 진단 능력의 정확도를 판단하기 위해 ROC 곡선을 이용한다.

ROC 곡선이란, 검사의 기준점(cutpoint) 변화에 따른 민감도와 1-특이도를 각각 Y 축과 X 축으로 하여 2차원으로 나타낸 그래프이다. 이는 어떤 검사의 최적 기준점을 찾거나 검사들의 진단 능력 비교에 주로 사용되는 통계 방법이다(Fawcett [2006]). 예를 들어, ROC 곡선은 질병 상태와 같은 독립적인 두 개의 집단을 구별할 때 사용할 수 있고, 유병률에 영향을 받지 않으며 진단기준에 관계 없이 전체적인 검사의 성능을 알 수 있는 장점이 있다. 이러한 면에서 ROC곡선은 진단검사 의학 분야에서 검사의 정확성을 파악하는 데 가장 많이 사용되고 있다(Begg [1991]). ROC 곡선을 기반으로 진단 도구의 정확도를 평가하는 경우, Youden [1950]이 제안한 Youden 계수가 가장 많이 사용되고 있다. Youden 계수(이하 J 로 표기)는 진양성률(True Positive Rate: TPR)에서

위양성률(False Positive Rate: FPR)을 $\text{1} - \text{민감도}$ 로 정의되며, 이를 최대화 하는 절단점 c 를 마커의 최적 절단점으로 추정한다. J 의 이론적인 범위는 -1에서 1 사이지만, 실제 사용하는 범위는 0에서 1 사이인데, 그 이유는 음수값은 실제 해석하는데 있어 의미가 없기 때문이다(Shan [2015]). J 의 값이 1이면 진단 검사법이 완벽함을 나타내고, 0이면 진단 검사법이 진단에 효과가 없음을 의미한다.

ROC곡선이 진단 도구의 정확도를 평가하는데 있어 가장 많이 사용되지만, 시간이 고정되어 있다고 가정하여 개인의 질병 상태를 평가하기 때문에 정보 손실의 위험이 있다. 왜냐하면 실제로 질병 상태와 마커값 모두 시간에 따라 변화하기 때문이다(Heagerty et al. [2000], Heagerty and Zheng [2005], Kamarudin et al. [2017]).

제 2 절 Time-dependent ROC 곡선을 이용한 절단점 추정

ROC 곡선은 마커 X 와 이분형 관심사건 발생여부 D 사이의 모든 가능한 c 값에 대해 추정한 민감도 $P(X > c|D = 1)$ 와, 특이도 $P(X \leq c|D = 0)$ 를 이용하여 그래프로 나타낸 것이다. D 가 시간에 종속되어 있는 경우, 민감도와 특이도는 각 c 에 대한 time-dependent 함수는 각각 $\text{Sens}(c, t) = P(X > c|D(t) = 1)$, $\text{Spec}(c, t) = P(X \leq c|D(t) = 0)$ 로 나타낼 수 있다. 이를 이용하여 각 시점 t 에서 ROC(t)로 정의할 수 있다.

한편, time-dependent ROC 곡선은 마커 사용 목적에 따라 다양한 $D(t)$ 를 고려할 수 있다. 본 논문에서는 $T \leq t$ 인 경우에 1, $T > t$ 인 경우에 0의 값을 가지는 $D_1(t)$ 와 $T = t$ 인 경우에 1, $T > t$ 인 경우에 0의 값을 가지는 $D_2(t)$ 를 기반으로 하는 민감도와 특이도를 이용하여 $D_1(t)$ 과 $D_2(t)$ 를 각각 정의하였다.

(I) $D_1(t)$ 의 경우 :

$$\text{Sens}^C(c, t) = P(X > c | T \leq t), \quad \text{Spec}^D(c, t) = P(X \leq c | T > t)$$

(II) $D_2(t)$ 의 경우 :

$$\text{Sens}^I(c, t) = P(X > c | T = t), \quad \text{Spec}^D(c, t) = P(X \leq c | T > t)$$

(I)은 치료 시점의 파악을 위해, 기준점에서 측정된 마커 값을 사용한 예후 정확도 평가 도구로써, (II)는 치료 결정을 내려야 하는 연구에서 기준점 또는 여러 시점에서 측정된 마커의 성능 평가의 도구로 적합하다(Heagerty and Zheng [2005], Bansal and Heagerty [2019], (Heagerty and Zheng [2005], Kamarudin et al. [2017]). Heagerty and Zheng [2005]는 (I)를 cumulative/dynamic(C/D), (II)를 incident/dynamic(I/D)으로도 정의하였다.

$i = 1, \dots, n$ 일 때, T_i 를 생존시간, C_i 를 중도절단 시간이라고 하면, 실제 관측자료는 $Y_i = \min(T_i, C_i)$ 로 나타낼 수 있으며, 사건 지시변수는 $\delta_i = I(T_i \leq C_i)$ 이다. 마커 X_i 로 나타낼 때, Choi [2020]는 $J(c, t)$ 의 추정량을 이용하여 time-dependent ROC 곡선을 단일 값으로 요약한 방법을 제시하였다.

$$\hat{c}^J = \underset{c}{\operatorname{argmax}} \frac{\sum_t \hat{J}(c, t) \times I(D(t) = 1)}{\sum_t I(D(t) = 1)}$$

여기서, $D(t) = 1$ 은 $\delta_i = 1$ 이고, $Y_i = t$ 일 때를 나타낸다. 이 추정량은 중도절단 되지 않은 관측된 생존시간에서 time-dependent ROC 곡선을 고려하고 각 시점별 $\hat{J}(c, t)$ 의 평균을 사용하고 있어, 중도절단된 자료의 정보를 충분히 활용하지 않았다.

본 논문에서는 관측된 생존시간의 범위를 균일하게 나누고, 균일하게 나눈 시점에서의 time-dependent ROC 곡선을 고려하여 각 시점 별로 적절한 가중치를 준 $\hat{J}(c, t)$ 의 가중평균을 최대로 하는 c 를 최적 절단점의 추정량으로 제안하였다.

$$\hat{c}^W = \operatorname{argmax}_c \frac{\sum_t \hat{J}(c, t) \times W(t)}{\sum_t W(t)}$$

가중치 $W(t)$ 로는 t 시점에서 생존시간 T 의 확률밀도함수값의 중도절단을 고려한 커널 평활 추정치를 사용하였다(Földes et al. [1981]).

$$W(t) = \sum_{i=1}^n K_b(t - Y_{(i)}) (\hat{S}_T(Y_{(i-1)}) - \hat{S}_T(Y_{(i)}))$$

이 때 $Y_{(i)}$ 는 Y_i 들의 i 번째 순서통계량이고 $\hat{S}_T(t)$ 는 생존시간 T 의 생존함수 $S_T(t)$ 의 Kaplan-Meier 추정량이다. 여기에서 $\hat{J}(c, t)$ 은 $J(c, t)$ 의 추정량이다. $J(c, t)$ 는 민감도와 특이도의 함수이므로 \hat{c}^J 과 \hat{c}^W 은 민감도와 특이도 추정량에 의존한다.

2.1 최근접 이웃(Nearest Neighbor: NN)방법

마커의 사용법에 따라 $D(t)$ 를 정의한 방법 중 $D_1(t)$ 인 경우 첫 번째 방법은 Heagerty et al. [2000]가 제시한 방법이다. Heagerty et al. [2000]는 시간

종속된 민감도와 특이도를 추정하기 위해 Kaplan-Meier를 추정량을 사용하였으나, 단조롭지 않고 0과 1사이의 값을 가지지 않는 단점이 있어 아래와 같은 방법을 제안하였다.

$$\text{Sens}(c, t) = \frac{P(X > c, T \leq t)}{P(T \leq t)} = \frac{1 - F_X(c) - S(c, t)}{1 - S_T(t)} \quad (2.1)$$

$$\text{Spec}(c, t) = \frac{P(X \leq c, T > t)}{P(T > t)} = \frac{S_T(t) - S(c, t)}{S_T(t)} = 1 - \frac{S(c, t)}{S_T(t)} \quad (2.2)$$

여기서 생존시간 T 의 생존함수는 $S_T(t)$, 마커의 누적분포 함수는 $F_X(c)$ 이며 $S(c, t)$ 는 (X, T) 의 이변량 생존함수를 $P(X > c, T > t)$ 로 나타내었다. Heagerty et al. [2000]은 (2.1)과 (2.2)에서 $S_T(t)$ 와 $S(c, t)$ 의 추정을 위해 Akritas [1994]가 제안한 이변량 생존함수 추정량인 최근접 이웃 추정량 (nearest neighbor estimator: NNE)의 사용을 제안하였다.

$$\hat{S}_{\lambda_n}(c, t) = \frac{1}{n} \sum_i \hat{S}_{\lambda_n}(t|X = X_i)I(X_i > c)$$

여기에서

$$\hat{S}_{\lambda_n}(t|X = X_i) = \prod_{s \in \tau_n, s \leq t} \left\{ 1 - \frac{\sum_j K_{\lambda_n}(X_j, X_i)I(Y_j = s)\delta_j}{\sum_j K_{\lambda_n}(X_j, X_i)I(Y_j \geq s)} \right\}$$

은 가중 Kaplan-Meier 추정량이고, $K_{\lambda_n}(X_j, X_i) = I(-\lambda_n < \hat{F}_X(X_j) - \hat{F}_X(X_i) < \lambda_n)$ 는 평활모수(smoothing parameter)가 λ_n 인 최근접 이웃 커널함수이다. X 의 분포함수 추정량은 $\hat{F}_X(c) = \sum_i I(X_i \leq c)/n$, T 의 생존함수 추정량은 $\hat{S}_{\lambda_n}(t) = \hat{S}_{\lambda_n}(-\infty, t)$ 으로, Heagerty et al. [2000]는 시간의 함수인

민감도와 특이도의 최근접 이웃 추정량을 다음과 같이 제안하였다.

$$\widehat{\text{Sens}}_N(c, t) = \frac{(1 - \widehat{F}_X(c)) - \widehat{S}_{\lambda_n}(c, t)}{1 - \widehat{S}_{\lambda_n}(t)}$$

$$\widehat{\text{Spec}}_N(c, t) = 1 - \frac{\widehat{S}_{\lambda_n}(c, t)}{\widehat{S}_{\lambda_n}(t)}$$

이 방법으로 추정된 $J(c, t)$ 를 사용하는 Choi [2020]가 제안한 최적 절단점 추정량과 본 논문에서 제안한 추정량을 각각 $\hat{c}^{JN}, \hat{c}^{WN}$ 으로 나타내고, 모의실험을 통해 비교하였다.

2.2 중도절단 가중치의 역확률 (Inverse Probability of Censoring Weighting: IPCW) 방법

$D_1(t)$ 의 두 번째 방법은 중도절단 가중치의 역확률 방법, IPCW이다. IPCW 방법은 Uno et al. [2007]가 제안하고 Hung and Chiang [2010]이 수정한 방법으로, 민감도 계산 시 가중치는 중도절단 되지 않을 확률의 추정치로 t 시점 전에 중도절단 되지 않은 개체에 중도절단 생존함수의 추정치의 역수로 만들어진다. 시간종속 민감도의 IPCW추정량은 아래와 같다.

$$\widehat{\text{Sens}}_I(c, t) = \frac{\sum_i I(X_i > c, T_i \leq t) \{\delta_i / n \widehat{S}_C(T_i)\}}{\sum_i I(T_i \leq t) \{\delta_i / n \widehat{S}_C(T_i)\}}$$

여기에서 $\widehat{S}_C(t)$ 는 중도절단시간 C 의 생존함수 $S_C(t)$ 의 Kaplan-Meier 추정량이다. Hung and Chiang [2010]은 특이도로 다음과 같은 단순(naive) 추정량을 사용하였다.

$$\widehat{\text{Spec}}_I(c, t) = \frac{\sum_i I(X_i \leq c, T_i > t)}{\sum_i I(T_i > t)}$$

이 방법은 단조로우며, 0과 1 사이의 값을 가진다.

위에서 정의한 민감도와 특이도의 추정량을 사용하여 Choi [2020]가 제안한 최적 절단점 추정량, 본 논문에서 제안한 최적 절단점 추정량을 각각 $\hat{c}^{JI}, \hat{c}^{WI}$ 로 나타내고 모의실험을 통하여 비교해보았다.

2.3 비례 위험 (Proportional Hazards: PH) 방법

$D(t)$ 이 $D_2(t)$ 인 경우, Heagerty and Zheng [2005]은 비례위험모형 $\lambda(t|X) = \lambda_0(t)\exp(X\gamma)$ 을 사용하여 민감도와 특이도를 추정하였다. 비례 가정을 완화하기 위해 비례위험 회귀계수 γ 를 시간에 따라 변하는 (time-varying) 계수로 모형화해서 평활 회귀를 통하여 $\hat{\gamma}(t)$ 로 추정한 후, 이를 이용하여 민감도를 추정하였다.

$$\widehat{\text{Sens}}_P(c, t) = \sum_i I(X_i > c) \pi_i(\hat{\gamma}(t), t)$$

$\pi_i(\gamma(t), t) = I(Y_i \geq t) \exp(X_i \gamma(t)) / W(t)$ 는 비례 위험 모델에서의 가중치이며, $W(t) = \sum_i I(Y_i \geq t) \exp(X_i \gamma(t))$ 이다. γ 가 시간에 따라 변하도록 허용하므로 실제로는 비례성을 만족하는 모형은 아니다. 특이도는 다음과 같이 경험적 (empirically)으로 추정할 수 있다.

$$\widehat{\text{Spec}}_P(c, t) = 1 - \sum_i I(X_i > c) \frac{\mathcal{R}_i^0(t)}{n_t}$$

$R_i^0(t)$ 는 사건이 발생하지 않은 개체여부, $n_t = \sum_k R_k^0(t)$ 를 나타낸다. $\widehat{\text{Sens}}_P(c, t)$ 와 $\widehat{\text{Spec}}_P(c, t)$ 를 사용하여 Choi [2020]의 제안 방법인 최적 절단점

추정량과 본 논문에서 제안한 최적 절단점 추정량을 $\hat{c}^{JP}, \hat{c}^{WP}$ 으로 표현하였으며, 모의실험을 통해 비교하였다.

제 3 장 모의실험

제 1 절 자료의 생성

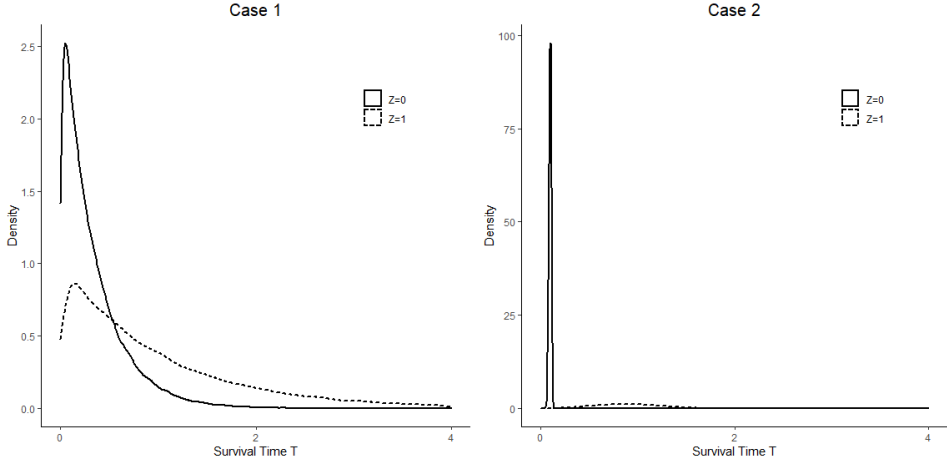
X (연속형 마커)는 평균이 0이고 분산이 4인 정규분포에서 생성하였다. c^* 가 실제 절단점일 때, 이진 변수 $Z = I(X \leq c^*)$ 를 생성하기 위해 절단점 c^* 를 0으로 설정하였다. 각각의 Z 값에 대한 생존 시간 T 값에 대해 Z 와 T 는 2가지 상황에서 형상모수(shape parameter) κ 와 척도모수(scale parameter) λ 를 가지고, 중도절단 시간 C 와 독립인 와이불(Weibull) 분포를 따른다. 와이불 분포의 확률밀도함수는 아래와 같다.

$$f_T(t; \kappa, \lambda) = \frac{\kappa}{\lambda} \left(\frac{t}{\lambda} \right)^{\kappa-1} e^{-(t/\lambda)^\kappa} I(t > 0)$$

표 3.1: The distribution of survival time T given $Z = z$

Case	$Z = 0$	$Z = 1$
	(κ_0, λ_0)	(κ_1, λ_1)
1	(1,3)	(1,1)
2	(1,3)	(10,1)

그림 3.1: Density plot for Weibull Distribution of T



T 와 독립인 중도절단 시간 C 는 $\text{uniform}(0, \tau)$ 를 따르며, 사건 지시 변수인 $\delta = I(T \leq C)$ 로 정의하였다. τ 값은 중도절단률을 제어하는 값이다.

제 2 절 모의실험 설계

모의실험을 위해 생성된 전체 표본은 각각 $N=50, 100, 300, 500$ 으로 설정하였고, 중도절단률은 20%, 40%, 60%로 설정하였다. 모든 분석은 통계소프트웨어인 R version 4.1.1을 이용하였다. 가중치 $W(t)$ 계산은 R 패키지 `survPresmooth` (López-de Ullibarri and Jácome [2013])를 이용하였다. 최근접 이웃 방법, 중도절단 가중치의 역확률방법, 비례위험방법에 대한 분석은 R 패키지 `survivalROC` (Heagerty et al. [2013]), `timeROC` (Blanche and Blanche [2019]), `risksetROC` (Heagerty et al. [2012])를 각각 이용하여 분석

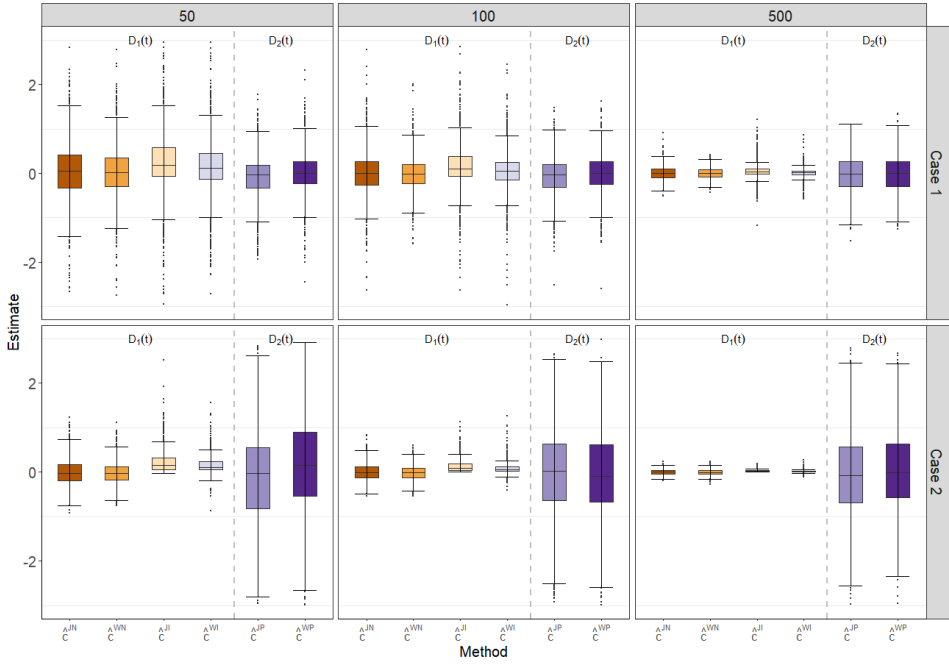
하였다. 1000번 반복 시행하여 각 추정치의 평균을 계산하였고, RMSE(Root Mean Square Error)를 이용하여 각 방법의 정확도를 평가하였다.

제 3 절 모의실험 결과

모의실험 결과는 그림 3.2와 부록 A.1로 정리하였다. 그림 3.2는 중도절단이 40%일 때, 각 case에 따라 $N=50, 100, 500$ 의 결과를 나타내었다. Case 1의 경우 $\kappa_0 = \kappa_1$ 이므로 비례위험가정을 만족하는 경우이고, case 2의 경우 $\kappa_0 \neq \kappa_1$ 이므로 비례위험가정을 만족하지 않는 경우이다. N 이 작을수록 이상치가 많이 존재 하여, 그림 3.2의 Y축의 범위 표시를 -2에서 2까지 나타내었다 (case 1에서 $N=50$ 일 때, $\hat{c}^{JI}, \hat{c}^{WI}$ 의 이상치는 약120개).

$D_1(t)$ 로 정의한 추정량들을 비교한 결과는 다음과 같다. 먼저, $D_1(t)$ 의 모든 추정량들을 표본의 크기에 따라 비교한 결과, 표본이 커짐에 따라 상자 그림의 크기가 점차 줄어드는 경향을 보였으며, 실제로 RMSE 값이 감소하는 경향을 보였다(부록 A.1). $D_1(t)$ 에서 Choi [2020]의 방법과, 본 논문에서 제시한 방법들을 비교한 결과, 그림 3.2에서 특히 표본의 수가 50일 경우, 이상치가 많아 \hat{c}^J 와 \hat{c}^W 가 비슷한 것처럼 보일 수 있지만(case 1에서 $N=50$ 인 경우, \hat{c}^{JN} : 9%, \hat{c}^{WN} : 3%, $\hat{c}^{JI}, \hat{c}^{WI}$: 12%), RMSE값을 확인 한 결과 \hat{c}^W 의 값이 더 작게 나타났다. 표본이 커지는 경우에도 \hat{c}^W 의 RMSE가 가장 작은 값을 나타내었으며, 그 중 가장 작은 값을 나타낸 추정량은 비례위험가정을 만족하는 case 1에서 \hat{c}^{WN} 가 가장 작은 값을 나타내었고, 표본이 커짐에 따라서 \hat{c}^{WN} 과 \hat{c}^{WP}

그림 3.2: Survival times in the two groups meet proportional hazards assumption(Case 1) and present moderate departures from proportional hazards assumption(Case 2) when censoring rate 40%



의 RMSE값이 비슷해지는 경향을 보였다. 비례위험가정을 만족하지 않는 case 2에서 \hat{c}^{WI} 이었다(부록 A.1).

$D_2(t)$ 의 추정량들을 비교한 결과는 다음과 같다. 표본이 작을 수록 이상치가 많아 추정량이 좋지 않은 것처럼 보이지만(case 1에서 표본의 수가 50인 경우, 이상치는 6%), RMSE값을 보면 \hat{c}^{WP} 의 값이 작은 결과를 나타내었으며, 전반적인 모의실험 환경에서 표본이 커짐에 따라, RMSE가 작았다(부록 A.1).

각 추정량들을 비교한 결과, case 1에서는 모든 상황에서 \hat{c}^{WP} 값이 작았고, case 2에서는 표본이 커질 수록 \hat{c}^{WP} 의 RMSE값이 줄어들었다(부록 A.1).

제 4 장 결론

연속형 변수의 이분화는 실제 임상 연구에서 해석이 쉽고 편리하기 때문에 일반적으로 사용하고 있다. 관심 있는 사건이 생존 자료일 때, 연속형 마커를 이분화 하는 최적의 절단점을 추정을 위해 가장 많이 사용되는 방법은 ROC 곡선을 기반으로 하는 방법이다. ROC곡선은 Youden index를 이용하여 연속형 마커를 이분화 하는 최적의 절단점 추정을 위해 사용된다. 그러나, 실제 질병 상태와 연속형 마커 값은 시간에 따라 변화하기 때문에 기준점(baseline)에서 개인의 질병 상태를 평가하여 얻어지는 ROC 곡선을 사용하는 방법은 최적의 절단점을 추정 시 정보 손실의 위험이 있다. 이를 대안으로 Heagerty et al. [2000]은 생존 자료 및 시간에 따라 변화하는 것을 고려한 time-dependent ROC 곡선을 이용하여 마커 최적의 절단점을 추정할 수 있으나, 이 또한 특정 시점에서의 time-dependent ROC 곡선만을 이용하는 한계점이 있다.

Choi [2020]은 time-dependent ROC 곡선의 중도절단 되지 않은 모든 생존시간에서 구한 Youden index들의 합이 최대가 되는 절단점을 마커의 최적의 절단점으로 추정하는 방법을 제안하여 주어진 정보를 충분히 활용하고자 하였다. 본 연구에서는 Choi [2020]의 방법을 수정하여 time-dependent ROC 곡선의 각 시점별 Youden index들의 가중합이 최대가 되는 값으로 최적 절단점을 추정하는 방법을 제안하였다.

본 논문에서 시행한 모의실험은 최적의 절단점을 찾기 위해 Choi [2020]가 제안한 추정량과 본 논문에서 제안한 추정량을 비교하였다. 모의실험은 각각 표본 크기, 중도절단률, 그리고 생존시간의 분포 설정을 다르게 지정하여 수행하였다. 각 방법들의 성능 평가는 RMSE를 통해 비교하였다.

모의실험 결과, 6가지 모든 추정량들은 N 이 작을 수록, 중도절단률이 커질 수록 RMSE값이 커짐을 확인하였다. 마커의 사용 목적에 따라 정의한 $D_1(t)$, $D_2(t)$ 의 결과를 각각 확인한 결과, $D_1(t)$ 의 경우 이상치가 많아 \hat{c}^J , \hat{c}^W 가 비슷한 것처럼 보였으나 실제 RMSE값을 확인하였을 때 \hat{c}^W 가 더 작은 결과를 나타내었다. 또한 각 추정량끼리 비교해보았을 때, 비례위험가정을 만족하는 경우 \hat{c}^{WN} 의 RMSE가 가장 작았으며 N 이 커질 수록 \hat{c}^{WI} 의 RMSE와 값이 비슷해지는 경향을 보였다. 비례위험가정을 만족하지 않는 경우 \hat{c}^{WI} 의 RMSE값이 가장 작음을 확인하였다. $D_2(t)$ 의 경우, RMSE값을 확인하였을 때 N 이 커질 수록 \hat{c}^{WP} 의 값이 더 작게 나타남을 확인하였다. $D_1(t)$ 와 $D_2(t)$ 는 마커의 사용법에 따라 각각 정의되었으므로, 추후 연구의 목적에 맞게 각 방법을 이용하면 될 것으로 보인다.

본 논문에서는 모의실험 환경을 일부만 설정하였으므로, 추후 연구에서는 더 많은 환경에서 비교를 해야할 것이다. 또한, 본 논문에서 제시한 방법인 \hat{c}^W 의 방법들과 기존에 많이 사용되고 있는 방법인 로그 순위 검정 등 다양한 방법들과 비교하는 것도 필요하며, 실제 자료에도 적용해보면 좋을 것으로 생각된다.

참고 문헌

- Akritis, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics*, pages 1299–1327.
- Bansal, A. and Heagerty, P. J. (2019). A comparison of landmark methods and time-dependent roc methods to evaluate the time-varying performance of prognostic markers for survival outcomes. *Diagnostic and prognostic research*, 3(1):1–13.
- Begg, C. B. (1991). Advances in statistical methodology for diagnostic medicine in the 1980's. *Statistics in medicine*, 10(12):1887–1895.
- Blanche, P. and Blanche, M. P. (2019). Package ‘timeroc’.
- Choi, S. (2020). *Determining optimal cutpoints in continuous biomarkers utilizing a time-dependent ROC curve*. PhD thesis, Yonsei University.
- Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.
- Földes, A., Rejtő, L., and Winter, B. (1981). Strong consistency properties

- of nonparametric estimators for randomly censored data, ii: Estimation of density and failure rate. *Periodica Mathematica Hungarica*, 12(1):15–29.
- Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344.
- Heagerty, P. J., Saha-Chaudhuri, P., and Saha-Chaudhuri, M. P. (2012). Package ‘risksetroc’.
- Heagerty, P. J., Saha-Chaudhuri, P., and Saha-Chaudhuri, M. P. (2013). Package ‘survivalroc’. *San Francisco: GitHub*.
- Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105.
- Hung, H. and Chiang, C.-t. (2010). Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data. *Scandinavian journal of statistics*, 37(4):664–679.
- Kamarudin, A. N., Cox, T., and Kolamunnage-Dona, R. (2017). Time-dependent roc curve analysis in medical research: current methods and applications. *BMC medical research methodology*, 17(1):1–19.
- Kim, J., Ng, H. K. T., and Kim, S. W. (2019). Assessing the optimal cutpoint

- for tumor size in patients with lung cancer based on linear rank statistics in a competing risks framework. *Yonsei medical journal*, 60(6):517–524.
- López-de Ullibarri, I. and Jácome, M. A. (2013). survpresmooth: an r package for presmoothed estimation in survival analysis. American Statistical Association.
- Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in medicine*, 25(1):127–141.
- Shan, G. (2015). Improved confidence intervals for the youden index. *PloS one*, 10(7):e0127272.
- Sima, C. S. and Gönen, M. (2013). Optimal cutpoint estimation with censored data. *Journal of Statistical Theory and Practice*, 7(2):345–359.
- Uno, H., Cai, T., Tian, L., and Wei, L.-J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537.
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

국문초록

Time-dependent ROC 곡선을 이용한 연속형 바이오마커의 최적의 절단점을 찾는 방법 비교: 모의 실험 연구

의학 연구에서 마커의 이분화는 해석이 쉽고 편리하기 때문에 많이 사용하고 있다. 가장 많이 사용되는 방법인 ROC 곡선은 Youden index를 이용하여 최적의 절단점 추정을 위해 사용된다. 그러나, 실제 질병 상태와 마커는 시간에 따라 변화 하므로 정보 손실 위험이 있다. 이를 대안으로 Heagerty et al. [2000]이 제안한 time-dependent ROC 곡선으로 마커의 최적 절단점을 추정할 수 있지만, 특정 시점에서 time-dependent ROC 곡선만 이용하는 한계가 있다. Choi [2020]은 time-dependent ROC 곡선의 중도절단 되지 않은 모든 생존시간에서 구한 Youden index의 합이 최대가 되는 값으로 마커의 최적 절단점을 추정하는 방법을 제안하였고, 본 논문에선 이를 수정하여 time-dependent ROC 곡선의 각 시점별 Youden index들의 가중합이 최대가 되는 값으로 마커의 최적 절단점을 추정하는 방법을 제안하였다. 모의실험을 통해 다양한 생존자료의 중도절단을 및 크기에 대하여 Choi [2020]가 제안한 방법과 본 논문에서 제안한 방법의 성능을 RMSE(Root Mean Square Error)를 통해 비교하였고, 본 논문에서 제안한 방법의 RMSE가 더 작은 결과를 나타내었다.

주요어 : 이분화, 예측 정확도, 시간 종속 ROC, Youden 계수

감사의 글

어느 새 2년간의 석사과정을 마치고 학위논문을 제출하게 되었습니다. 그동안 저에게 많은 도움을 주신 분들께 감사인사를 전하려 합니다. 먼저 너무 부족했던 저에게 올바른 학문의 길로 이끌게 해주신 최혜미 교수님께 정말 감사하다는 말씀을 전하고 싶습니다. 교수님의 지도 덕분에 연구를 할 때 어떤 방향으로 해야할 지 많이 배웠습니다. 바쁘신 중에도 저의 심사를 맡아주신 양성준 교수님과 최규빈 교수님께도 감사의 말씀을 전합니다. 많은 지식을 쌓게 도와주신 김대경 교수님, 한경수 교수님, 조성일 교수님, 마주칠 때마다 항상 격려와 조언을 아끼지 않으신 안정용 교수님과 정성석 교수님 정말 감사드립니다. 지난 2년간 제가 더 발전할 수 있도록 많은 격려와 아낌없는 지원을 해주신 전북대학교병원 임상약리센터 김민걸 센터장님을 비롯한 모든 센터직원분들, 특히 자기일처럼 나서주었던 임상통계팀 선생님들께 감사하다는 말씀 전하고 싶습니다. 대학원 생활하면서 많은 의지가 되었던 은진언니와 바울이, 선 언니, 무한한 사랑으로 옆에서 묵묵히 응원해준 아버지 그리고 진선, 용우, 어디서든 지켜보고 있을 어머니에게 고맙고 사랑한다는 말을 전하고 싶습니다. 마지막으로 무슨일이 있어도 항상 제 편이 되어준 남편 이동환님께도 정말 고맙고 사랑한다고 전하고 싶습니다. 감사한 마음을 보답하기 위해 앞으로도 더 발전해 나가는 사람이 될 수 있도록 노력하겠습니다.

2022년 1월

조 진 실

부록 A 모의실험 결과 표

⌘ A.1: Survival times in the two groups meet proportional hazards assumption(Case 1) and present moderate departures from proportional hazards assumption(Case 2) when censoring rate 40%

Case	N	Statistics	Method					
			$D_1(t)$			$D_2(t)$		
			\hat{C}^{JN}	\hat{C}^{WN}	\hat{C}^{JI}	\hat{C}^{WI}	\hat{C}^{JP}	\hat{C}^{JP}
1	50	Mean	0.0503	0.0346	0.2435	0.1426	-0.0827	0.0021
		RMSE	0.7695	0.6800	0.8794	0.7442	0.5139	0.5077
	100	Mean	0.0148	-0.0169	0.1558	0.0520	-0.0646	-0.0129
		RMSE	0.5284	0.3902	0.6175	0.5014	0.4579	0.4513
	500	Mean	0.0042	-0.0025	0.0461	0.0065	-0.0217	-0.0092
		RMSE	0.1618	0.1229	0.1849	0.1243	0.4187	0.4116
2	50	Mean	0.0002	-0.0234	0.2361	0.1631	-0.1127	0.1169
		RMSE	0.3985	0.2577	0.3568	0.2508	1.1778	1.1830
	100	Mean	-0.0035	-0.0205	0.1347	0.0865	-0.0356	-0.0949
		RMSE	0.1882	0.1647	0.2028	0.1537	1.0483	0.9948
	500	Mean	-0.0014	-0.0070	0.0254	0.0168	-0.0652	0.0131
		RMSE	0.0648	0.0634	0.0380	0.0323	0.9638	0.9091