The Korean Journal of Applied Statistics
2021, Vol. 34, No. 2, 141–152

DOI: https://doi.org/10.5351/KJAS.2021.34.2.141
Print ISSN1225-066X / Online ISSN 2383-5818

Undecided inference using the difference of AUCs

Chong Sun Hong[1,a], Hae Rin Na[a]

[a]Department of Statistics, Sungkyunkwan University

Abstract

A new statistical model needs additional variables in order to re-evaluate the undecided inference. Then the MNAR assumption is required, since the probabilities for the positivity of the indeterminant and the determinant is calculated differently. In this study, since two statistical models have a hierarchical relationship, we determine the undecided inference under the MNAR assumption using the confidence interval of the difference between two AUCs. Among many methods of estimating the confidence interval of the AUC difference, it is found that four kinds of methods show excellent performance through simulations. And based on these methods, we propose a variable selection method that are useful for the undecided inference using logistic regression models.

Keywords: hierarchy, indeterminant, logistic regression, missing, variable selection

1. 서론

신용평가와 의학통계분야 등에서 많이 사용하는 receiver operating characteristic (ROC) 곡선은 이진분류모형(binary classification model)의 성능(performance)을 탐색하는 유용한 시각적인 방법이며, 모든 분류점(threshold, cut-off point)에 의한 실제 양성(positive)을 양성이라고 정확하게 판단하는 true positive rate (TPR)과 실제 음성(negative)을 양성이라고 잘못 판단하는 false positive rate (FPR)의 변화를 나타내는 곡선이다 (Egan, 1975; Metz, 1978; Provost와 Fawcett, 2001; Vuk와 Curk, 2006). ROC 곡선에서 이진분류모형 또는 분류자의 판별력을 측정하는 대표적인 측도로 ROC 곡선 아래 면적을 나타내는 AUC(area under the ROC curve)가 있다 (Hanley와 McNeil, 1982; Swets, 1988; Centor, 1991; Bradley, 1997; Engelmann 등, 2003; Joseph, 2005; Hong 등, 2013).

일반적으로 사용하는 분류모형에서는 로지스틱 모형, 프로빗 모형, maximum rank correlation (MRC) 추정 모형 등의 방법을 이용하여 여러 개의 확률변수를 선형결합한 선형스코어(linear score)를 사용한다 (Su와 Liu, 1993; Pepe와 Thompson, 2000; Pepe 등, 2006; Hong과 Won, 2016; Hong 등, 2019). 선형스코어 모형에 새로운 변수를 추가한 후 두 모형의 차이에 대한 통계적 검정방법으로는 왈드(Wald) 검정과 가능도비(likelihood ratio) 검정방법 등이 존재한다. 새로운 모형은 기존 모형보다 유의하면, 새로운 모형의 판별력도 역시 유의한 증가하므로 AUC 차이에 대한 신뢰구간을 추정한다 (Pepe 등, 2013).

AUC 차이에 대한 신뢰구간을 추정하는 연구 중에서 비모수적 방법인 맨휘트니 U-통계량(Mann-Whitney U-statistic)을 이용하는 Hanley와 McNeil (1983), DeLong 등 (1988), 그리고 Bandos 등 (2007)의 추정 방법이

[1] Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-Ro, Jongno-Gu, Seoul 03063, Korea. E-mail: cshong@skku.edu

Published 30 April 2021 / journal homepage: http://kjas.or.kr
© 2021 The Korean Statistical Society. All rights reserved.

존재한다. Hanley와 McNeil (1983)은 두 모형의 양성과 음성에 대한 켄달의 순위상관(kendall's rank correlation)을 이용하여 AUC 차이의 신뢰구간을 추정하였고, DeLong 등 (1988)은 맨휘트니 U-통계량을 기반으로 일반화된 U-통계량(generalized U-statistic)의 성질을 이용해 두 AUC 차이의 점근적인 분포를 추정하여 두 AUC 차이의 신뢰구간을 제안하였으며, Bandos 등 (2007)은 모든 가능한 붓스트랩(bootstrap) 표본을 고려하여 두 AUC 차이 신뢰구간의 정확한 붓스트랩 분산 추정량을 구하는 방법을 제안하였다. 그리고 Li 등 (2008)은 이변량 정규분포일때의 AUC 추정량을 이용하여 두 AUC의 차에 일반화된 축량(generalized pivotal quantities) 개념을 적용하여 몬테카를로 알고리즘(Monte Carlo algorithm)을 통해 두 AUC 차의 신뢰구간을 추정하는 방법을 제안하였다. 또한 Heller 등 (2017)은 평활된 경험적 AUC (smoothed empirical AUC)를 이용하며, 이 AUC를 최대로 하는 MRC 추정량을 모형의 계수로 이용하여 계층(hierarchical, nested) 관계에 있는 두 모형의 AUC 차이에 대한 점근적인 분포를 추정하는 신뢰구간 방법과 점근적인 분포의 분산과 AUC 차이의 강한 선형관계를 제거하기 위하여 제곱근 변환한 신뢰구간 방법을 제안하였다.

신용평가, 의학통계 등 다양한 통계적 모형의 분석과정에서 이진분류가 결정이 어려워 판단이 보류하는 경우가 자주 발생한다. 이런 경우를 판단이 보류된 미결정자(undecided)라고 한다. 예를 들어 신용평가에선 승인점 전략에 의해 차주의 대출 여부를 결정하기 어려운 중간 점수대의 차주나 의학에선 환자의 질병 유무를 판단하기 어려운 환자에게 추가적인 정보를 필요로 하는 경우이다 (Hong과 Jung, 2011a, 2011b). Feelders (2000)와 Hand (2001)는 미결정자 추론 또는 거절자 추론(reject inference)을 결측자료(missing data) 문제로 간주하여 결측값의 유형에 따라 미결정자를 구분하였다. 일부 개체에 대해 평가가 이루어지지 않아 판단이 보류된 미결정자는 missing at random (MAR) 방법으로 접근하며, 특정 구간에서 평가가 이루어지지 않아 추가적인 정보를 이용하여 재평가하기 위해 결정이 보류된 미결정자는 missing not at random (MNAR) 방법으로 간주한다 (Hong과 Jung, 2011a, 2011b). MAR 가정의 경우 미결정자와 결정자의 분포가 같지만, MNAR 가정의 경우 미결정자의 분포와 결정자의 분포는 달라야 하므로 미결정자 추론에 도움이 되는 변수를 추가한 모형을 미결정자의 모형으로 사용한다. Hong과 Jung (2011b)은 로지스틱 회귀모형을, Hong과 Jung (2011a)은 이변량 프로빗 모형을 이용하여 MAR 가정과 MNAR 가정에 대한 미결정자 추론방법을 제안하였다.

본 논문에서는 Hong과 Jung (2011b)의 MNAR 가정하에서 미결정자 추론을 바탕으로 새로운 정보가 포함된 확률변수를 추가한 미결정자 추론모형이 기존 모형보다 통계적으로 유의함을 기존 모형과 새로운 모형에 대하여 가설검정과 더불어 AUC 차이 신뢰구간 추정방법을 이용하여 미결정자에서 판별력이 높은 변수를 선택하는 방법을 제안한다. 그리고 다변량 정규분포 가정을 통해 하나의 독립변수를 가진 모형과 두 개의 독립변수를 가진 계층 모형을 설정하여 계층모형의 가능도비 검정을 통해 판별력의 증가를 확인하고 다양한 AUC 차이 신뢰구간 추정방법들이 이를 잘 반영하는지를 탐색하여 AUC 차이 신뢰구간 추정방법들의 성능을 비교 연구한다.

본 논문의 구성은 다음과 같다. 2절에서는 미결정자 추론에서 기존 모형과 새로운 변수가 추가된 모형의 AUC 차이 신뢰구간을 응용하는 방법을 제안하고, 3절에서는 여섯 종류의 AUC 차이의 신뢰구간 추정 방법들의 성능을 정규분포 가정하에 비교하여 미결정자 추론에 적합한 방법에 대하여 연구한다. 4절에서는 실증예제를 통해 미결정자 추론에서 AUC 차이 신뢰구간을 구하여 추가된 확률변수들의 유의성을 판단하면서 변수선택하는 방법을 보여주며, 마지막 5절에서는 본 논문의 결론을 서술한다.

2. AUC 차이 신뢰구간을 이용한 미결정자 추론

자료의 관찰된 확률표본을 k차원 확률변수 행렬 $\mathbf{X}$라고 표기하자. 이 자료를 활용한 판별모형으로부터 얻은 선형스코어 확률변수는 $\mathbf{X}$의 함수인 $L(\mathbf{X})$이고, 이에 대한 양성유무를 판단하는 확률변수 Y는 양성($Y = 1$)과 음성($Y = 0$)으로 구성되며, 미결정자 유무를 판단하는 보조변수 A는 미결정자이면 $A = 1$, 결정자이면 $A = 0$

이다.

미결정자 추론의 MAR 가정하에서 $\mathbf{X}$의 조건부 결정자는 Y에 의존하지 않고, 미결정자 Y의 분포는 결정자 Y의 분포와 같으므로 미결정자의 모형은 결정자로부터의 모형과 같은 모형을 사용한다.

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x}, A = 1) = P(Y = 1|\mathbf{X} = \mathbf{x}, A = 0).$$

MAR 가정의 경우에는 결정자를 이용하여 모형을 추정하고, 이를 미결정자 모형으로 사용하여 미결정자 자료를 적용해 미결정자의 미래상태를 판단한다.

Hong과 Jung (2011a)은 미결정자 그룹($A = 1$)의 양성과 음성에 대한 판단이 결정자 그룹($A = 0$)의 변수 $\mathbf{X}$의 영향력과 다르거나 $\mathbf{X}$이외에 추가적으로 영향을 주는 요소가 존재하여 미결정자와 결정자의 양성률은 다르기 때문에 MAR 가정이 아닌 MNAR 가정을 사용했다. 즉, $\mathbf{X} = \mathbf{x}$의 조건부 결정자가 Y에 의존하고, 미결정자 Y의 분포는 결정자 Y의 분포와 다르므로

$$P(Y = 1|\mathbf{X} = \mathbf{x}, A = 1) \neq P(Y = 1|\mathbf{X} = \mathbf{x}, A = 0)$$

이며, 이 확률은 $P(Y = 1|\mathbf{X} = \mathbf{x})$와 동일하지 않으므로 MNAR 가정에서 미결정자 추론은 결정자로부터의 모형과 다른 모형을 사용한다.

Hong과 Jung (2011a, 2011b)은 MNAR 가정에서 결정자 모형에 사용한 변수 $\mathbf{X} = (X_1, \ldots, X_k)$외에 확률변수 $\mathbf{X}^+ = (X_{k+1}, \ldots, X_p)$를 추가한 확률변수 $\mathbf{X}^* = (\mathbf{X}, \mathbf{X}^+) = (X_1, \ldots, X_k, X_{k+1}, \ldots, X_p)$를 이용하여 로지스틱 회귀모형 혹은 이변량 프로빗 모형을 통해 미결정자의 분류모형을 추정하였다. 본 연구에서는 로지스틱 회귀모형을 적용하여 MNAR 가정에서 두 AUC 차이 신뢰구간을 이용한 미결정자 추론을 설명한다. 기존의 변수 $\mathbf{X}$를 이용한 미결정자 추론과 확률변수 $\mathbf{X}^*$를 이용한 미결정자 추론에서 추가된 $\mathbf{X}^+$의 선택이 통계적으로 유의했는지를 판단하는 방법으로 두 모형의 AUC 차이에 대한 신뢰구간을 추정한다.

먼저, 선형스코어 확률변수 $L(\mathbf{X})$를 로지스틱 회귀 모형 (혹은 이변량 프로빗 모형)으로 추정한다. 즉, $L(\mathbf{X}) = \text{logit}\, P(Y = 1|\mathbf{X}) = \hat{\beta}_0 + \hat{\beta}'\mathbf{X} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$. 이때 MNAR 가정의 미결정자 보조변수는 $A = I(c_1 \leq L(\mathbf{X}) \leq c_2)$이며, $\mathbf{X}$와 $\mathbf{X}^*$를 사용한 미결정자 추론 모형은 각각 다음과 같이 구한다.

$$\log\left(\frac{\hat{P}}{1-\hat{P}}\right) = \hat{\beta}_0 + \hat{\beta}'\mathbf{X}, \qquad \log\left(\frac{\hat{P}^*}{1-\hat{P}^*}\right) = \hat{\beta}_0 + \hat{\beta}'\mathbf{X}^*, \tag{2.1}$$

여기서 $\hat{P} = \hat{P}(Y = 1|\mathbf{X} = \mathbf{x}, A = 1)$과 $\hat{P}^* = \hat{P}(Y = 1|\mathbf{X}^* = \mathbf{x}^*, A = 1)$는 미결정자의 양성률이다.

$\mathbf{X}$를 양성($Y = 1$)인 자료의 $\mathbf{X^1}$과 음성($Y = 0$)인 자료인 $\mathbf{X^0}$으로 분할하고, $\mathbf{X}^*$를 $Y = 1$인 자료의 $\mathbf{X}^{*\mathbf{1}}$과 $Y = 0$인 자료인 $\mathbf{X}^{*\mathbf{0}}$으로 분할하면 다음과 같이 나타낼 수 있다.

$$\mathbf{X} = \begin{pmatrix} \mathbf{X^1} \\ \mathbf{X^0} \end{pmatrix}, \quad \mathbf{X}^* = \begin{pmatrix} \mathbf{X}^{*\mathbf{1}} \\ \mathbf{X}^{*\mathbf{0}} \end{pmatrix} = \begin{pmatrix} \mathbf{X^1} & \mathbf{X}^{+\mathbf{1}} \\ \mathbf{X^0} & \mathbf{X}^{+\mathbf{0}} \end{pmatrix}. \tag{2.2}$$

식 (2.2)의 표현을 통해 $\mathbf{X}$와 $\mathbf{X}^*$를 이용한 미결정자에 대한 AUC 통계량을 각각 AUC_1과 AUC_2라 하고 다음과 같이 나타난다.

$$\begin{aligned} \text{AUC}_1 &= P\left(\hat{\beta}_0 + \hat{\beta}'\mathbf{X^1} > \hat{\beta}_0 + \hat{\beta}'\mathbf{X^0}|A = 1\right), \\ \text{AUC}_2 &= P\left(\hat{\beta}_0 + \hat{\beta}'\mathbf{X}^{*\mathbf{1}} > \hat{\beta}_0 + \hat{\beta}'\mathbf{X}^{*\mathbf{0}}|A = 1\right). \end{aligned} \tag{2.3}$$

$\mathbf{X}^*$는 $\mathbf{X}$를 포함하는 계층적 관계에 있으며 $\mathbf{X}^*$의 분류모형이 통계적으로 유의하면 $\mathbf{X}^*$를 사용한 AUC_2가 $\mathbf{X}$를 사용한 AUC_1보다 크다 (Pepe 등, 2013). 그러므로 두 AUC의 차이 $\delta = \text{AUC}_2 - \text{AUC}_1$의 신뢰구간을 추정하여 $\mathbf{X}^*$를 사용했을 때가 $\mathbf{X}$를 사용했을 때보다 미결정자의 AUC가 유의하게 증가했는지 확인하고, 증가가

유의하면 변수 $\mathbf{X}^+$를 그대로 사용하고 증가가 유의하지 않으면 변수 $\mathbf{X}^+$가 아닌 다른 변수를 고려한다. 그러므로 미결정자를 추론하기 위한 방법으로 기존의 모형에 새로운 확률변수 $\mathbf{X}^+$의 유의성을 판단하기 위하여 다음과 같은 방법을 제안한다.

Proposal

미결정자를 추론하기 위하여 기존 모형에 확률변수 $\mathbf{X}^+$를 추가한 새로운 모형과의 비교 검정결과가 통계적으로 유의할 뿐만아니라 두 모형에 대한 AUC 차의 신뢰구간에 0이 포함하지 않으면, 추가된 확률변수 $\mathbf{X}^+$들이 유의하다고 판단되어 모형에 추가하는 것을 추천한다.

다음 절에서는 여섯 종류의 AUC 차이 신뢰구간 추정방법들의 성능을 파악하기 위하여, 확률변수들의 분포를 다변량 정규분포로 설정하여 비교한다.

3. AUC 차이 신뢰구간 방법들의 비교

Bandos 등 (2007), Li 등 (2008), 그리고 Kim (2010)은 양성 집단과 음성 집단이 일변량인 대응표본의 AUC 차이 신뢰구간의 포함확률(coverage probability)과 기대길이(expected length)를 비교하였으며, Li의 방법의 포함확률은 명목수준 95%를 상회하지만 DeLong과 Bandos의 방법은 명목수준을 하회하며, 기대길이는 대체적으로 Li의 방법이 DeLong과 Bandos의 방법보다 크다는 것을 밝혔다. 하지만 새롭게 추정된 미결정자 모형과 기존 미결정자 모형은 계층 관계에 있으므로, 본 연구는 양성과 음성의 집단이 일변량이 아닌 두 개의 계층 모형에 대한 AUC의 변화 유무를 가능도비 검정을 통해 확인하고 AUC 차이 신뢰구간 추정 방법들을 바탕으로 AUC의 변화 유무를 잘 반영하는지 탐색하여 비교한다.

식 (2.2)에서 $\mathbf{X} = (X_1^1\ X_1^0)'$와 $\mathbf{X}^* = (\mathbf{X}^{*\mathbf{1}}\ \mathbf{X}^{*\mathbf{0}})'$로 설정하고, $\mathbf{X}^{*\mathbf{1}}$과 $\mathbf{X}^{*\mathbf{0}}$를 각각 다음과 같은 이변량 정규분포(bivariate normal; BVN)로 설정한다.

$$\mathbf{X}^{*\mathbf{1}\prime} = \begin{pmatrix} X_1^1 \\ X_2^1 \end{pmatrix} \sim \text{BVN}\left(\begin{pmatrix} 0.5 \\ \mu \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \quad \mathbf{X}^{*\mathbf{0}\prime} = \begin{pmatrix} X_1^0 \\ X_2^0 \end{pmatrix} \sim \text{BVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

양성집단의 두 번째 변수 X_2^1의 평균 μ를 0.4부터 0.65까지 0.5 간격으로 변화시키며, 상관계수 ρ를 -0.8부터 0.8까지 0.1씩 증가시킨다. 모든 변수에 대한 표본을 각각 30개씩 추출하여 식 (2.1)과 같이 기존 변수 $\mathbf{X} = X_1$과 모든 변수 $\mathbf{X}^* = (X_1, X_2)$를 사용한 로지스틱 회귀 모형을 구하며 식 (2.3)을 이용하여 두 모형의 AUC 통계량을 구한다. 두 AUC 차이의 신뢰구간을 Hanley와 McNeil (1983), DeLong 등 (1988), Bandos 등 (2007), Li 등 (2008), Heller 등 (2017)의 제곱근 변환 전 방법과 제곱근 변환 방법을 이용하여 총 1,000번을 독립적으로 실행해 평균 신뢰구간을 구한다. 그리고 귀무가설을 $H_0 : L(\mathbf{X}) = L(\mathbf{X}^*)$으로 설정한 가능도비 검정을 실시하여 유의수준 0.05에서 귀무가설을 기각하면 AUC의 변화가 있다고 판단한다. 따라서 δ의 신뢰구간에 0의 포함 유무를 가능도비 검정의 귀무가설 기각여부와 비교한다.

Table 1과 Table 2에 각 μ와 ρ에 대한 가능도비 검정의 p-값(p-value)을 정리하고 p-값이 0.05 이하인 경우는 굵게 표시했다. AUC 차이 신뢰구간에 0이 포함되면 'O', 포함되지 않으면 'X'로 표현하며, Hanley와 McNeil (1983), DeLong 등 (1988), Bandos 등 (2007), Li 등 (2008)의 신뢰구간 추정 방법을 약자로 각각 HM, DL, BA, LI로 나타내고, Heller 등 (2017)의 제곱근 변환 전 방법과 제곱근 변환 방법을 약자로 각각 HE1, HE2로 나타내었다.

AUC 차이 신뢰구간 방법들의 성능을 비교하기 위해 판별력의 증가를 잘 반영하는지를 탐색하여 가능도비 검정의 귀무가설 기각 여부와 비교하였다. 새로운 모형이 통계적으로 유의하면 판별력의 증가가 일어나며 이를 가능도비 검정의 p-값을 통해 알 수 있다. p-값이 0.05보다 작은 경우 AUC 차이 신뢰구간에 0을 포함하지

Table 1: 95% confidence interval of the difference of AUCs

μ_d	ρ	p-value	δ	HM	DL	BA	δ	LI	δ	HE1	HE2
	−0.8	**0.0000**	**0.2105**	**X**	**X**	**X**	**0.2073**	**X**	0.2036	**X**	**X**
	−0.7	**0.0001**	**0.1669**	**X**	**X**	**X**	**0.1634**	**X**	0.1616	**X**	**X**
	−0.6	**0.0008**	**0.1456**	**O**	**X**	**X**	**0.1425**	**X**	0.1428	**X**	**X**
	−0.5	**0.0034**	**0.1291**	**O**	**O**	**O**	**0.1274**	**O**	0.1281	**O**	**X**
	−0.4	**0.0094**	**0.1222**	**O**	**O**	**O**	**0.1196**	**O**	0.1219	**O**	**X**
	−0.3	**0.0179**	**0.1127**	**O**	**O**	**O**	**0.1119**	**O**	0.1155	**O**	**X**
	−0.2	**0.0334**	**0.1119**	**O**	**O**	**O**	**0.1109**	**O**	0.1162	**O**	**X**
	−0.1	**0.0497**	**0.1083**	**O**	**O**	**O**	**0.1056**	**O**	0.1149	**O**	**X**
0.4	0.0	0.0766	0.1044	O	O	O	0.1032	O	0.1149	O	X
	0.1	0.0926	0.0988	O	O	O	0.0980	O	0.1136	O	X
	0.2	0.1374	0.1009	O	O	O	0.0996	O	0.1224	X	X
	0.3	0.1600	0.0975	O	O	O	0.0968	O	0.1233	X	X
	0.4	0.1884	0.0964	O	O	O	0.0948	O	0.1298	X	X
	0.5	0.2094	0.0925	O	O	O	0.0916	O	0.1334	X	X
	0.6	0.2564	0.0923	O	O	O	0.0909	O	0.1345	X	X
	0.7	0.2961	0.0883	O	O	O	0.0887	O	0.1242	X	X
	0.8	0.2966	0.0896	O	O	O	0.0886	O	0.1036	O	X
	−0.8	**0.0000**	**0.2222**	**X**	**X**	**X**	**0.2183**	**X**	**0.2145**	**X**	**X**
	−0.7	**0.0000**	**0.1784**	**X**	**X**	**X**	**0.1753**	**X**	**0.1732**	**X**	**X**
	−0.6	**0.0004**	**0.1524**	**X**	**X**	**X**	**0.1496**	**X**	**0.1495**	**X**	**X**
	−0.5	**0.0017**	**0.1383**	**O**	**O**	**O**	**0.1363**	**X**	**0.1368**	**X**	**X**
	−0.4	**0.0053**	**0.1223**	**O**	**O**	**O**	**0.1215**	**O**	**0.1225**	**O**	**X**
	−0.3	**0.0106**	**0.1170**	**O**	**O**	**O**	**0.1163**	**O**	**0.1196**	**O**	**X**
	−0.2	**0.0171**	**0.1143**	**O**	**O**	**O**	**0.1123**	**O**	**0.1187**	**O**	**X**
	−0.1	**0.0314**	**0.1122**	**O**	**O**	**O**	**0.1100**	**O**	**0.1191**	**O**	**X**
0.45	0.0	**0.0496**	**0.1085**	**O**	**O**	**O**	**0.1058**	**O**	**0.1183**	**O**	**X**
	0.1	0.0632	0.1029	O	O	O	0.1017	O	0.1185	O	X
	0.2	0.1024	0.1002	O	O	O	0.0984	O	0.1208	X	X
	0.3	0.1331	0.0987	O	O	O	0.0980	O	0.1260	X	X
	0.4	0.1569	0.0988	O	O	O	0.0977	O	0.1351	X	X
	0.5	0.1933	0.0952	O	O	O	0.0947	O	0.1405	X	X
	0.6	0.2159	0.0942	O	O	O	0.0930	O	0.1402	X	X
	0.7	0.2418	0.0917	O	O	O	0.0914	O	0.1369	X	X
	0.8	0.2981	0.0877	O	O	O	0.0874	O	0.1154	O	X
	−0.8	**0.0000**	**0.2307**	**X**	**X**	**X**	**0.2276**	**X**	**0.2229**	**X**	**X**
	−0.7	**0.0000**	**0.1912**	**X**	**X**	**X**	**0.1883**	**X**	**0.1860**	**X**	**X**
	−0.6	**0.0002**	**0.1585**	**X**	**X**	**X**	**0.1570**	**X**	**0.1553**	**X**	**X**
	−0.5	**0.0009**	**0.1433**	**O**	**X**	**X**	**0.1412**	**X**	**0.1420**	**X**	**X**
	−0.4	**0.0029**	**0.1311**	**O**	**O**	**O**	**0.1285**	**O**	**0.1314**	**O**	**X**
	−0.3	**0.0060**	**0.1191**	**O**	**O**	**O**	**0.1187**	**O**	**0.1220**	**O**	**X**
	−0.2	**0.0130**	**0.1164**	**O**	**O**	**O**	**0.1151**	**O**	**0.1218**	**O**	**X**
	−0.1	**0.0210**	**0.1135**	**O**	**O**	**O**	**0.1122**	**O**	**0.1210**	**O**	**X**
0.5	0.0	**0.0326**	**0.1116**	**O**	**O**	**O**	**0.1111**	**O**	**0.1236**	**O**	**X**
	0.1	0.0534	0.1079	O	O	O	0.1062	O	0.1248	X	X
	0.2	0.0723	0.1072	O	O	O	0.1062	O	0.1308	X	X
	0.3	0.0785	0.1016	O	O	O	0.1002	O	0.1310	X	X
	0.4	0.1109	0.1001	O	O	O	0.0994	O	0.1374	X	X
	0.5	0.1335	0.1009	O	O	O	0.1007	O	0.1434	X	X
	0.6	0.1627	0.0978	O	O	O	0.0975	O	0.1456	X	X
	0.7	0.1821	0.0967	O	O	O	0.0956	O	0.1391	X	X
	0.8	0.2271	0.0945	O	O	O	0.0921	O	0.1247	X	X

Table 2: 95% confidence interval of the difference of AUCs

μ_d	δ	p-value	δ	HM	DL	BA	δ	LI	δ	HE1	HE2
	−0.8	**0.0000**	**0.2473**	**X**	**X**	**X**	**0.2443**	**X**	**0.2399**	**X**	**X**
	−0.7	**0.0000**	**0.2005**	**X**	**X**	**X**	**0.1977**	**X**	**0.1946**	**X**	**X**
	−0.6	**0.0001**	**0.1705**	**X**	**X**	**X**	**0.1684**	**X**	**0.1678**	**X**	**X**
	−0.5	**0.0005**	**0.1497**	**O**	**X**	**X**	**0.1470**	**X**	**0.1481**	**X**	**X**
	−0.4	**0.0015**	**0.1359**	**O**	**O**	**O**	**0.1342**	**X**	**0.1364**	**X**	**X**
	−0.3	**0.0038**	**0.1257**	**O**	**O**	**O**	**0.1243**	**O**	**0.1283**	**O**	**X**
	−0.2	**0.0075**	**0.1233**	**O**	**O**	**O**	**0.1218**	**O**	**0.1280**	**O**	**X**
	−0.1	**0.0141**	**0.1174**	**O**	**O**	**O**	**0.1151**	**O**	**0.1249**	**O**	**X**
0.55	0.0	**0.0223**	**0.1157**	**O**	**O**	**O**	**0.1125**	**O**	**0.1267**	**X**	**X**
	0.1	**0.0340**	**0.1101**	**O**	**O**	**O**	**0.1086**	**O**	**0.1266**	**X**	**X**
	0.2	**0.0410**	**0.1060**	**O**	**O**	**O**	**0.1046**	**O**	**0.1296**	**X**	**X**
	0.3	0.0583	0.1054	O	O	O	0.1039	O	0.1371	X	X
	0.4	0.0820	0.1030	O	O	O	0.1017	O	0.1406	X	X
	0.5	0.0943	0.1015	O	O	O	0.1008	O	0.1488	X	X
	0.6	0.1210	0.0983	O	O	O	0.0979	O	0.1475	X	X
	0.7	0.1533	0.0968	O	O	O	0.0957	O	0.1416	X	X
	0.8	0.1542	0.0976	O	O	O	0.0969	O	0.1223	O	X
	−0.8	**0.0000**	**0.2538**	**X**	**X**	**X**	**0.2504**	**X**	**0.2459**	**X**	**X**
	−0.7	**0.0000**	**0.2096**	**X**	**X**	**X**	**0.2061**	**X**	**0.2040**	**X**	**X**
	−0.6	**0.0000**	**0.1770**	**X**	**X**	**X**	**0.1748**	**X**	**0.1737**	**X**	**X**
	−0.5	**0.0002**	**0.1576**	**X**	**X**	**X**	**0.1555**	**X**	**0.1558**	**X**	**X**
	−0.4	**0.0007**	**0.1423**	**O**	**X**	**X**	**0.1399**	**X**	**0.1423**	**X**	**X**
	−0.3	**0.0023**	**0.1330**	**O**	**O**	**O**	**0.1311**	**O**	**0.1360**	**X**	**X**
	−0.2	**0.0057**	**0.1263**	**O**	**O**	**O**	**0.1247**	**O**	**0.1317**	**X**	**X**
	−0.1	**0.0083**	**0.1184**	**O**	**O**	**O**	**0.1172**	**O**	**0.1265**	**O**	**X**
0.6	0 .0	**0.0155**	**0.1146**	**O**	**O**	**O**	**0.1133**	**O**	**0.1272**	**X**	**X**
	0.1	**0.0225**	**0.1133**	**O**	**O**	**O**	**0.1119**	**O**	**0.1309**	**X**	**X**
	0.2	**0.0285**	**0.1107**	**O**	**O**	**O**	**0.1096**	**O**	**0.1352**	**X**	**X**
	0.3	**0.0431**	**0.1114**	**O**	**O**	**O**	**0.1095**	**O**	**0.1431**	**X**	**X**
	0.4	0.0534	0.1046	O	O	O	0.1044	O	0.1458	X	X
	0.5	0.0673	0.1044	O	O	O	0.1043	O	0.1532	X	X
	0.6	0.0761	0.1003	O	O	O	0.0994	O	0.1510	X	X
	0.7	0.1018	0.1018	O	O	O	0.1008	O	0.1502	X	X
	0.8	0.1095	0.1005	O	O	O	0.0996	O	0.1298	X	X
	−0.8	**0.0000**	**0.2685**	**X**	**X**	**X**	**0.2643**	**X**	**0.2609**	**X**	**X**
	−0.7	**0.0000**	**0.2178**	**X**	**X**	**X**	**0.2143**	**X**	**0.2123**	**X**	**X**
	−0.6	**0.0000**	**0.1924**	**X**	**X**	**X**	**0.1881**	**X**	**0.1888**	**X**	**X**
	−0.5	**0.0001**	**0.1634**	**X**	**X**	**X**	**0.1611**	**X**	**0.1615**	**X**	**X**
	−0.4	**0.0004**	**0.1495**	**O**	**X**	**X**	**0.1470**	**X**	**0.1500**	**X**	**X**
	−0.3	**0.0011**	**0.1398**	**O**	**O**	**O**	**0.1374**	**X**	**0.1420**	**X**	**X**
	−0.2	**0.0028**	**0.1301**	**O**	**O**	**O**	**0.1281**	**O**	**0.1356**	**X**	**X**
	−0.1	**0.0057**	**0.1256**	**O**	**O**	**O**	**0.1236**	**O**	**0.1343**	**X**	**X**
0.65	0.0	**0.0078**	**0.1214**	**O**	**O**	**O**	**0.1192**	**O**	**0.1338**	**X**	**X**
	0.1	**0.0153**	**0.1142**	**O**	**O**	**O**	**0.1135**	**O**	**0.1337**	**X**	**X**
	0.2	**0.0215**	**0.1140**	**O**	**O**	**O**	**0.1123**	**O**	**0.1404**	**X**	**X**
	0.3	**0.0282**	**0.1113**	**O**	**O**	**O**	**0.1107**	**O**	**0.1448**	**X**	**X**
	0.4	**0.0397**	**0.1092**	**O**	**O**	**O**	**0.1074**	**O**	**0.1495**	**X**	**X**
	0.5	**0.0434**	**0.1054**	**O**	**O**	**O**	**0.1050**	**O**	**0.1577**	**X**	**X**
	0.6	**0.0441**	**0.1063**	**O**	**O**	**O**	**0.1049**	**O**	**0.1580**	**X**	**X**
	0.7	0.0572	0.1028	O	O	O	0.1029	O	0.1511	X	X
	0.8	0.0671	0.1065	O	O	O	0.1052	O	0.1297	X	X

않은 'X'가 많을 수록, p-값이 0.05보다 큰 경우 AUC 차이 신뢰구간에 0을 포함한 'O'가 많을 수록 판별력의 증가를 잘 반영하는 방법이라고 할 수 있다.

Table 3: Results of six methods

p-value		HM	DL	BA	LI	HE1	HE2
$p \geq 0.05$	X	0	0	0	0	32	38
	O	38	38	38	**38**	6	0
	Total			38			
$p < 0.05$	X	19	24	24	**27**	45	64
	O	45	40	40	37	19	0
	Total			64			

Table 4: Data information

Variable	Variable information	Characteristic
Y	diagnosis	binomial (0,1)
X_1	texture_worst	continuous
X_2	radius_se	
X_3	symmetry_worst	
X_4	concave.points_worst	
X_5	smoothness_worst	
X_6	compactness_mean	
X_7	perimeter_worst	

Table 1과 Table 2를 바탕으로 p-값이 0.05보다 큰 경우와 작은 경우에 대한 'O'와 'X'의 개수를 Table 3에 정리하였다. Table 3의 p-값이 0.05보다 큰 경우를 보면 HM, DL, BA, LI 방법이 모두 'X'와 'O'의 개수가 각각 0개, 38개로 동일하지만, HE1과 HE2 방법은 'X'의 개수가 각각 32개, 38개로 앞의 네 가지 방법보다 많으며, 'O'의 개수가 각각 6개, 0개로 앞의 네 가지 방법보다 적으므로 HE1과 HE2 방법은 가장 성능이 좋지 않다고 할 수 있다. HE1과 HE2 방법을 제외한 나머지 네 가지의 방법들에 대해 p-값이 0.05보다 작은 경우 'X'와 'O'의 개수를 비교해보면, LI 방법이 'X'가 27개로 가장 많고 'O'가 37개로 가장 적기 때문에 LI의 방법이 두 모형에 차이가 있을 때 AUC 차이의 신뢰구간에서 0을 포함하지 않는 것을 다른 방법들에 비해 잘 반영하는 방법이라고 할 수 있다. 그러므로 AUC 차이의 신뢰구간을 추정하는 다양한 방법 중 LI의 방법이 가장 성능이 좋고, 그다음 DL과 BA 방법, 그리고 마지막으로 HM 방법이 좋다고 할 수 있다. HE1과 HE2 방법은 분산이 불안정하여 AUC 차이 신뢰구간을 추정하는 것에 오류가 발생하는 것을 확인하였다.

본 연구의 모의실험 결과를 통해 성능이 좋지 않은 HE1과 HE2 방법을 제외한 나머지 네 가지 방법(HM, DL, BA, LI)을 MNAR 가정의 미결정자 추론에서 AUC 차이 신뢰구간을 추정할 때 사용한다.

4. 미결정자 추론 실증 예제

미결정자 추론에 네 가지 방법(HM, DL, BA, LI)을 사용한 AUC 차이 신뢰구간 추정을 실제 자료에 적용하여 분석한다. Yang 등 (2019)에서 사용한 'Wisconsin breast cancer data (diagnostic)'을 이용하여 MNAR 가정의 미결정자에서 AUC 차이 신뢰구간을 통해 미결정자 추론에 도움이 되는 변수를 선택하는 방법을 살펴본다. 자료는 30개의 독립변수와 총 569개의 유방암 진단 결과로 212개의 양성진단, 357개의 음성진단으로 구성되어있으며, 양성진단이면 $Y = 1$로 판단하고 음성진단일 때 $Y = 0$으로 판단한다. 전진선택법(forward selection)을 통해 변수 7개가 선택되었으며 자료의 구성은 아래 Table 4와 같다. 이 자료는 `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnositc)`에서 사용할 수 있다.

기존의 변수를 $\mathbf{X} = (X_1, X_2, X_3, X_4)$, 새롭게 추가한 변수를 $\mathbf{X}^+ = (X_5, X_6, X_7)$으로 설정하였으며, 선형스코어 확률변수 $L(\mathbf{X})$는 로지스틱 회귀 모형을 이용해 추정하여 분류점 c를 전체자료에서의 양성율 $P(Y = 1|\mathbf{X}) = 0.3729$로 설정하고, 이 절단점 c를 기준으로 전체의 하위 12.5%와 상위 12.5%를 각각 미결정자 보조변수 $A = I(c_1 = 0.0437 \leq L(\mathbf{X}) \leq c_2 = 0.9773)$의 c_1과 c_2으로 설정하여 c_1과 c_2 사이에 존재하는 전체자료의 25%

Table 5: 95% Confidence interval of the difference of AUCs

Method	AUC_1	AUC_2	δ	95% CI
HM				(0.0537,0.1643)
DL	0.8598	0.9688	0.1090	(0.0549,0.1632)
BA				(0.0549,0.1632)
LI	0.8820	0.9904	0.1084	(0.0635,0.1634)

Table 6: Confusion matrices for the undecided group data

Original		Actual	
		Positive	Negative
Pred.	Positive	54	17
	Negative	14	58

MNAR		Actual	
		Positive	Negative
Pred.	Positive	63	9
	Negative	5	66

Table 7: Confusion matrices for the overall data

Original		Actual	
		Positive	Negative
Pred.	Positive	198	18
	Negative	14	339

MNAR		Actual	
		Positive	Negative
Pred.	Positive	207	10
	Negative	5	347

를 MNAR 가정에서의 미결정자로 설정하면 미결정자 그룹의 자료는 양성진단이 68개, 음성진단이 75개인 총 143개의 자료로 구성된다.

이때 기존의 변수 $\mathbf{X} = (X_1, X_2, X_3, X_4)$에 대한 식 (2.1)의 왼쪽 식을 이용한 미결정자 모형은 다음과 같다.

$$\log\left(\frac{\hat{P}}{1-\hat{P}}\right) = -15.3306 + 10.2384X_1 + 28.0016X_2 + 4.2676X_3 + 16.7427X_4,$$

여기서 $\hat{P} = P(Y = 1|\mathbf{X} = \mathbf{x}, A = 1)$.

MNAR 가정하의 미결정자 추론을 위해 $\mathbf{X}^+$를 추가한 $\mathbf{X}^* = (X_1, \ldots, X_7)$에 대한 식 (2.1)의 오른쪽 식을 이용한 미결정자 모형은 다음과 같다.

$$\begin{aligned}\log\left(\frac{\hat{P}^*}{1-\hat{P}^*}\right) = &- 27.5765 + 12.2841X_1 + 36.1833X_2 + 5.4564X_3 + 16.0472X_4 \\ &+ 10.6694X_5 - 13.0696X_6 + 36.6126X_7,\end{aligned}$$

여기서 $\hat{P}^* = P(Y = 1|\mathbf{X}^* = \mathbf{x}^*, A = 1)$.

$\mathbf{X}$와 $\mathbf{X}^*$를 이용한 각각의 미결정자 자료에 대한 양성률을 계산하고, 이를 바탕으로 미결정자에서의 두 AUC 즉, 식 (2.2)에서의 AUC_2와 AUC_1의 차이($\delta = AUC_2 - AUC_1$)에 대한 신뢰구간을 네 가지 방법(HM, DL, BA, LI)으로 추정하여 $\mathbf{X}$에 $\mathbf{X}^+$를 추가했을 때 판별력이 더 좋아졌는지를 판단한다.

미결정자 AUC 차이 신뢰구간을 추정한 결과는 Table 5에 정리하였다. Table 5에서 MNAR 가정하에서 미결정자 그룹의 AUC_2와 기존 변수를 사용했을 때 미결정자 그룹의 AUC_1의 차이에 대한 95% 신뢰구간을 보면 네 가지 방법 모두 0을 포함하지 않고, 하한 또한 HM의 방법의 경우 0.0537, DL과 BA의 방법은 0.0549, LI의 방법은 0.0635로 0에 근사하지 않는 값을 가진다. 이 결과를 통해 새로운 변수 $\mathbf{X}^+ = (X_5, X_6, X_7)$를 추가하면 $\mathbf{X} = (X_1, X_2, X_3, X_4)$만 이용했을 때보다 미결정자 AUC가 유의하게 증가하므로 MNAR 가정의 미결정자 추론에 도움이 되는 변수라는 것을 알 수 있다.

원자료 $\mathbf{X}$에 $\mathbf{X}^+$가 추가된 $\mathbf{X}^*$를 사용한 MNAR 가정에서의 미결정자에 대한 혼동행렬을 Table 6에 나타내고, 원자료와 MNAR 가정에서의 전체에 대한 혼동행렬은 Table 7에 정리하였다. Table 6을 보면 미결정자 그룹에 대한 원자료와 미결정자 추론 후의 오분류율은 각각 21.68%, 11.89%이고, $\mathbf{X}^*$를 사용한 미결정자 추론의 오분류율이 원자료에서의 오분류율보다 약 11.89%만큼 감소했다. 그래서 Table 7를 보면 전체에 대한 원자료와 미결정자 추론 후의 오분류율은 각각 5.62%, 2.64%로 미결정자 추론을 한 후의 오분류율이 원자료보다 2.98% 감소했다는 것을 확인할 수 있다.

실증예제에서 미결정자 추론에 AUC 차이 신뢰구간을 이용하여 $\mathbf{X}^*$를 사용했을 때가 $\mathbf{X}$를 사용했을 때보다 미결정자의 AUC가 유의하게 증가했으므로 $\mathbf{X}^+$가 미결정자 추론에 도움을 주는 변수임을 쉽게 판단할 수 있다. 그리고 Table 6과 Table 7의 혼동행렬을 통해서도 미결정자 추론에서 $\mathbf{X}^*$를 사용하면 $\mathbf{X}$를 사용하는 것보다 오분류율이 더 감소하는 것을 확인할 수 있어 $\mathbf{X}^+$가 미결정자 추론에 도움을 주는 변수라는 것을 알 수 있다.

5. 결론

미결정자 추론을 결측자료 문제로 간주할 때, 기존 변수의 영향력이 미결정자 그룹과 결정자 그룹에서 다르거나 결정자 그룹에서 기존 변수 이외에 추가적으로 영향을 주는 변수가 존재하므로 미결정자와 결정자의 양성률이 다르게 추정되는 MNAR 가정이 필요하다. 본 연구에서는 MNAR 가정하의 미결정자 추론에서 AUC 차이의 신뢰구간 추정방법을 이용하여 미결정자 분류모형에 추가할 변수를 선택하는 방법을 제안한다. 새롭게 추정된 미결정자 모형에서의 AUC와 기존 미결정자 모형에서의 AUC의 차이에 대한 신뢰구간으로 새롭게 추정된 미결정자 모형이 기존 미결정자 모형보다 판별력이 큰지를 판단하며, 이를 바탕으로 추가된 변수가 미결정자 추론에 도움을 주는 변수인지를 확인할 수 있다.

새롭게 추정된 미결정자 모형과 기존 미결정자 모형은 계층 관계에 있으므로, 이변량 정규분포를 이용해 계층 모형을 설정하고 두 계층 모형의 가능도비 검정을 통해 AUC 차이가 실제로 일어났는지를 판단하여 여러 가지 AUC 차이 신뢰구간 추정방법에 대한 성능을 비교하였다. 그 결과 AUC가 차이가 없을 것이라고 예상될 때, Hanley와 McNeil (1983), DeLong 등 (1988), Bandos 등 (2007), Li 등 (2008)의 방법의 성능은 동일하지만, Heller 등 (2017)의 두 가지 방법은 대부분 AUC 차이가 있다고 판단하므로 AUC 차이를 제대로 반영하지 못한다는 것을 확인하였다. 그리고 Heller 등 (2017)의 방법들을 제외한 나머지 네 가지의 방법 중에서 Li 등 (2008) 방법이 가장 성능이 좋으며 DeLong 등 (1988)과 Bandos 등 (2007)의 방법은 동일한 성능을 가지며, 그다음으로 Hanley와 McNeil (1983)의 방법이 좋다는 것을 발견하였다. 따라서 성능이 좋은 네 가지 방법(Hanley와 McNeil, 1983; DeLong 등, 1988; Bandos 등, 2007; Li 등, 2008)을 미결정자 추론에 사용하였다.

실증예제를 통해 로지스틱 회귀를 이용한 MNAR 가정의 미결정자 추론에서 AUC 차이 신뢰구간을 바탕으로 미결정자 추론에 도움이 되는 변수의 예를 설명했다. 기존 변수 $\mathbf{X}$와 추가할 변수 $\mathbf{X}^+$를 설정하여 미결정자 추론 후의 미결정자 AUC와 기존의 미결정자 AUC의 차이가 유의하므로 본 논문에서 설정한 변수들을 미결정자 추론에 사용하였다. 이 변수들을 이용하여 미결정자 추론 후의 혼동행렬과 원자료의 혼동행렬의 오분류율을 비교하였을 때 미결정자 추론 후의 오분류율이 낮은 것을 확인하였다. 그러므로 본 연구에서 제안한 AUC 차이 신뢰구간을 이용한 미결정자 추론방법은 MNAR 가정에서 미결정자에 영향이 있는 변수를 선택하는 방법으로 활용할 수 있다.

References

Bandos, A. I., Rockette, H. E., and Gur, D. (2007). Exact bootstrap variances of the area under ROC curve. *Communications in Statistics—Theory and Methods*, **36**, 2443–2461.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition*, **30**, 1145–1159.

Centor, R. M. (1991). Signal detectability: the use of ROC curves and their analyses. *Medical decision making*, **11**, 102–106.

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics*, **44**, 837–845.

Egan, J. P. (1975). *Signal Detection Theory and ROC-Analysis*, Academic Press.

Engelmann, B., Hayden, E., and Tasche, D. (2003). Testing rating accuracy, *Risk*, **16**, 82–86.

Feelders, A. J. (2000). Credit scoring and reject inference with mixture models, *International Journal of Intelligent System in Accounting*, **8**, 271–279.

Hand, D. J. (2001). Reject inference in credit operations, *Handbook of Credit Scoring*, 225–240.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, **143**, 29–36.

Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases, *Radiology*, **148**, 839–843.

Heller, G., Seshan, V. E., Moskowitz, C. S., and Gönen, M. (2017). Inference for the difference in the area under the ROC curve derived from nested binary regression models, *Biostatistics*, **18**, 260–274.

Hong, C. S. and Jung, M. H. (2011a). Undecided inference using bivariate probit models, *Journal of the Korean Data and Information Science Society*, **22**, 1017–1028.

Hong, C. S. and Jung, M. S. (2011b). Undecided inference using logistic regression for credit evaluation. *Journal of the Korean Data and Information Science Society*, **22**, 149–157.

Hong, C. S. and Won, C. H. (2016). Parameter estimation for the imbalanced credit scoring data using AUC maximization. *The Korean Journal of Applied Statistics*, **29**, 309–319.

Hong, C. S., Jeon, H. S., and Shin, H. S. (2019). Threshold interval for linear combination scores maximizing the partial AUC and VUS, *The Korean Data and Information Science Society*, **30**, 759–770.

Hong, C. S., Jung, E. S., and Jung, D. G. (2013). Standard criterion of VUS for ROC surface, *The Korean Journal of Applied Statistics*, **26**, 977–985.

Joseph, M. P. (2005). A PD validation framework for Basel II internal ratings-based systems, Credit Scoring and Credit Control IV.

Kim, H. Y. (2010). A comparison of the interval estimations for the difference in paired areas under the ROC curves, *Communications for Statistical Applications and Methods*, **17**, 275–292.

Li, C. R., Liao, C. T., and Liu, J. P. (2008). On the exact interval estimation for the difference in paired areas under the ROC curves, *Statistics in Medicine*, **27**, 224–242.

Metz, C. E. (1978). Basic principles of ROC analysis, *In Seminars in Nuclear Medicine*, **8**, 283–298.

Pepe, M. S., Cai, T., and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve, *Biometrics*, **62**, 221–229.

Pepe, M. S., Kerr, K. F., Longton, G., and Wang, Z. (2013). Testing for improvement in prediction model performance, *Statistics in Medicine*, **32**, 1467–1482.

Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics*, **1**, 123–140.

Provost, F. and Fawcett, T. (2001). Robust classification for imprecise environments, *Machine Learning*, **42**, 203–231.

Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers, *Journal of the American Statistical Association*, **88**, 1350–1355.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems, *Science*, **240**, 1285–1293.

Vuk, M. and Curk, T. (2006). ROC curve, lift chart and calibration plot, *Metodoloski Zvezki*, **3**, 89.

Yang, H., Lu, K., Lyu, X., and Hu, F. (2019). Two-way partial AUC and its properties, *Statistical Methods in Medical Research*, **28**, 184–195.

Received November 9, 2020; Revised December 7, 2020; Accepted January 5, 2021

AUC 차이를 이용한 미결정자 추론방법

홍종선[1,a], 나해린[a]

[a]성균관대학교 통계학과

요 약

미결정자 추론을 재평가하기 위해 기존 변수에 새로운 변수들을 추가하는 통계 모형이 필요하다. 미결정자와 결정자의 양성률은 다르게 계산되기 때문에 MNAR 가정이 필요하다. 본 연구에서는 두 통계적 모형이 계층 관계를 가지고 있으므로, 두 AUC 차이의 신뢰구간을 이용하여 MNAR 가정하에서 미결정자를 추론한다. AUC 차이 신뢰구간의 추정방법 중에서 모의실험을 통하여 네 종류의 방법의 성능이 우수함을 발견하였다. 그리고 네 종류의 방법을 바탕으로 로지스틱 회귀를 이용한 미결정자 추론에 도움이 되는 변수를 선택하는 방법을 제안한다.

주요용어: 결측, 계층, 로지스틱회귀, 미결정, 변수선택

[1]교신저자: (03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: cshong@skku.edu