The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.2.309
29(2), 309–319

Parameter estimation for the imbalanced credit scoring data using AUC maximization

C. S. Hong[a,1] · C. H. Won[a]

[a]Department of Statistics, Sungkyunkwan University

(Received November 2, 2015; Revised December 28, 2015; Accepted January 5, 2016)

Abstract

For binary classification models, we consider a risk score that is a function of linear scores and estimate the coefficients of the linear scores. There are two estimation methods: one is to obtain MLEs using logistic models and the other is to estimate by maximizing AUC. AUC approach estimates are better than MLEs when using logistic models under a general situation which does not support logistic assumptions. This paper considers imbalanced data that contains a smaller number of observations in the default class than those in the non-default for credit assessment models; consequently, the AUC approach is applied to imbalanced data. Various logit link functions are used as a link function to generate imbalanced data. It is found that predicted coefficients obtained by the AUC approach are equivalent to (or better) than those from logistic models for low default probability - imbalanced data.

Keywords: discrimination, link, risk, ROC, threshold

1. 서론

신용평가모형의 판별력은 부도와 정상의 차주들을 사전적으로 구별하는 능력을 의미하며, 이러한 신용평가모형의 판별력을 측정하는 기준에는 ROC(receiver operating characteristic), CAP(cumulative accuracy profile), GINI계수 등이 있다. ROC 곡선은 의학진단이나 신용평가에서 모형의 성능(performance)을 탐색하는 유용한 방법으로 모형의 정분류율과 오분류율의 변화를 시각적으로 나타내기 위해서 이용되어 왔으며, 특히 신용평가 분야와 같이 사전에 불량 거래자를 정확하게 판단해야하는 모형의 판별력에 대한 시각적인 방법으로 확장되었다 (Egan, 1975; Swets, 1988; Swets 등, 2000; Sobehart와 Keenan, 2001; Engelmann 등, 2003). ROC 곡선의 특성에 관한 설명과 응용에 관련된 정보는 Fawcett (2003), Provost와 Fawcett (2001), Hong과 Choi (2009), Hong 등 (2010)에서 발견할 수 있다.

분류를 위한 확률변수 X를 스코어(score) 변수라 하자. Y의 원소 0은 환자나 차주의 정상상태(good, non-default), 1은 부도상태(bad, default)로 정의하면, 분류점으로부터의 이항 결과를 $Y = \{0, 1\}$으로 나타낼 수 있다. Pepe 등 (2005)은 기존의 한 개의 스코어 변수로 이루어진 분류모형에서 여러 개의 변수들의 선형결합으로 이루어진 스코어 함수를 고려하였다. 선형 스코어(linear score)를 다음과 같이 정

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

의하였고, ROC 곡선의 불변성 성질(invariance property)을 기반으로 절편이 없으며 $\beta_1 = 1$로 설정하였다.

$$L_\beta(X) = X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

X_i^1는 $Y = 1$인 $i = 1, \ldots, n^1$개의 자료이고, X_j^2는 $Y = 0$인 $j = 1, \ldots, n^2$개인 자료라 하면, 스코어 변수를 대체한 선형 스코어를 이용하여 민감도와 1-특이도를 다음과 같이 정의하여 ROC 곡선은 가능한 모든 분류점 c에 대한 민감도와 1-특이도의 집합으로 표현된다.

$$\begin{aligned} \text{sens}(c) &= P\left(L_\beta\left(X_i^1\right) \geq c\right), \\ 1 - \text{spec}(c) &= P\left(L_\beta\left(X_j^2\right) \geq c\right), \\ \text{ROC}(c) &= \{(1 - \text{spec}(c), \text{sens}(c)),\ c \in (-\infty, \infty)\}. \end{aligned}$$

ROC 곡선의 평가기준 측도로 ROC 곡선의 아래의 면적을 계산한 AUC(area under the ROC curve)를 사용한다. AUC 값은 0.5와 1사이에 존재하며 1에 가까울수록 분류모형에 대한 판별력이 높다고 할 수 있다. Hosmer (2000)와 Josephe (2005)는 AUC 값의 크기로 모형의 판별력을 판단하는 기준들을 제안하였다.

Pepe (2003)는 리스크(risk) 스코어를 선형 스코어의 함수로 연결함수 $g(\cdot)$을 이용하여 다음과 같이 정의하고

$$P(Y = 1|X) = g(L_\beta(X)). \tag{1.1}$$

연결함수 $g(\cdot)$가 단조증가함수일 때, 네이만-피어슨 정리(Neyman-Pearson lemma)와 ROC 곡선의 성질에 의하여 기각영역 $L_\beta(X) > c$에서 최적의 ROC 곡선을 가진다는 것을 보였다. 즉 $L_\beta(X) > c$는 다른 어떤 선형 스코어 함수보다 최적의 ROC 곡선을 가진다. Bamber (1975)는 경험적 AUC 통계량을 Mann-Whitney 통계량으로도 표현하였고, Pepe 등 (2005)은 식 (1.2)와 같은 경험적 AUC 통계량을 목적함수로 설정하여 AUC를 최대화하는 β를 추정하였다.

$$\widehat{\text{AUC}}(b) = \frac{1}{n^1 n^2} \sum_{i=1}^{n^1} \sum_{j=1}^{n^2} \left\{ I\left[L_b\left(X_i^1\right) > L_b\left(X_j^2\right)\right] + 0.5 I\left[L_b\left(X_i^1\right) = L_b\left(X_j^2\right)\right] \right\}. \tag{1.2}$$

AUC 통계량을 목적함수로 하여 모수 β를 추정하는 방법을 $\widehat{\beta}^{AUC} = \arg\max \widehat{\text{AUC}}(b)$로 표현하고 이 방법을 AUC 접근방법이라고 한다. 식 (1.2)은 계단형의 이산형식이므로 Newton-Raphson 알고리즘을 대신하여 본 연구에서는 Cavanagh와 Sherman (1998)이 적용한 NM 알고리즘 (Nelder와 Mead, 1965)을 SAS로 구현하여 AUC 접근방법의 모수를 추정하였다.

추정량 $\widehat{\beta}$은 MRC(maximum rank correlation) 추정량의 특별한 경우이며 (Han, 1987), 적어도 하나의 설명변수가 연속형인 일반화 선형모형 하에서 일치성과 정규근사성(asymptotic normality)을 가진다 (Sherman, 1993). Pepe 등 (2005)은 AUC 접근방법의 추정량의 좋은 점은 자료가 로지스틱 가정하에서 또는 가정이 맞지 않는 일반적인 상황에 대하여도 여전히 로지스틱모형의 추정량보다 동등하거나 더 좋은 것이라 주장하였다. 이와 관련된 연구는 확장되어 Kraus (2014)는 로지스틱 가정이 맞지 않는 상황을 로짓(logit), 프로빗(probit), complementary log-log 등의 다섯 가지 연결함수 $g(\cdot)$들을 이용하여 추정량을 구하고 비교 토론하였다.

ROC 곡선은 다항 범주의 판별 결과로 확장되어 세 범주 분류에서의 ROC 곡면(surface)과 네 가지 이상의 다항범주 분류에서의 ROC 다면체(manifold) 그리고 AUC 통계량에 대응하는 VUS(the volum

under the ROC surface)와 HUM(the hyper-volume under the ROC manifold)에 대하여 정의되고 다양한 판별기준이 제안 되었다 (Scurfield, 1996; Mossman, 1999; Dreiseitl 등, 2000; Heckerling, 2001; Fawcett, 2003; Nakas와 Yiannoutsos, 2004; Patel과 Markey, 2005; Zou 등, 2007; Li와 Fine, 2008; Wandishin과 Mullen, 2009; Nakas 등, 2010; Hong 등, 2013; Hong과 Jung, 2013; Hong과 Jung, 2014; Hong과 Zhi Qiang, 2014). Hong과 Cho (2015a, 2015b)는 VUS와 HUM을 조건부 Mann-Whitney 통계량과 Wilcoxon 순위합 통계량으로 표현하고 가설검정을 수행하였다. 또한 Hong 등 (2015)은 VUS와 HUM에 Pepe (2005)의 방법을 확장하여 다음과 같은 모수 추정방법인 $\widehat{\beta}^{VUS} = \arg\max \widehat{\mathrm{VUS}}(b)$와 $\widehat{\beta}^{HUM} = \arg\max \widehat{\mathrm{HUM}}(b)$을 VUS, HUM 접근방법으로 제안하고 AUC 접근방법과 유사하게 로지스틱 가정이 맞지 않는 상황에서도 로지스틱 모형의 추정량과 동등하거나 좋은 결과를 보이는 것을 확인하였다.

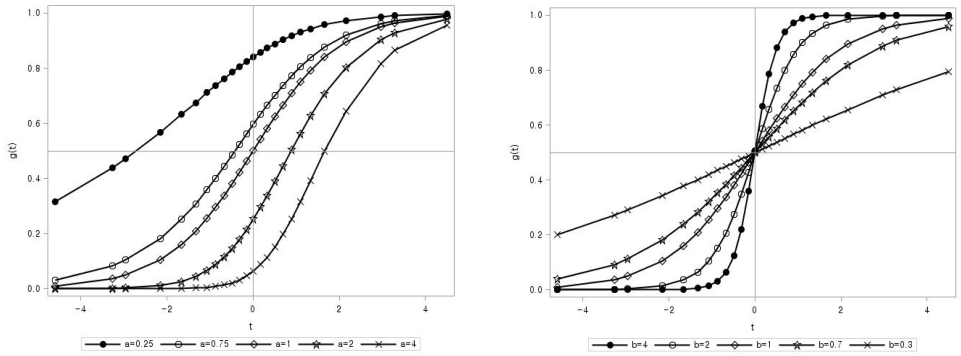
본 연구에서는 일반적인 신용평가 관점에서 흔하게 발생하는 상황인 낮은 부도율의 상황을 고려한다. 신 BIS 자기자본규제제도의 도입으로 인하여 신용평가시스템의 구축 및 운영 이슈 사항으로 부도 수가 적은 포트폴리오에 대하여 유형을 제시되었고, 일 년 또는 그 이하의 짧은 기간 동안의 평가시스템의 판별력 추정에 있어서 실제로 일반적인 많은 포트폴리오에는 신용등급이 있는 기업의 수가 1,000을 넘지 않고, 그 결과 부도의 수가 희박한 것은 당연하다고 볼 수 있다 (Tasche, 2009). 부도수가 희박한 불균형 자료에 대하여 모수 추정의 문제가 발생하고 로지스틱 모형의 판별력에 대한 다양한 문제가 제기 되었고 (Allison, 2008; Calabrese와 Osmetti, 2011; Brown과 Mues, 2012), 이에 대하여 불균형 자료에서 선형모형의 판별력을 최대화시키는 모수 추정에 관한 연구를 한다.

본 논문의 구성은 다음과 같다. 논문의 2절에서는 불균형 자료의 모수 추정의 문제에 대하여 설명하고 낮은 부도율 상황을 가정하기 위한 변형된 로짓 함수를 설명한다. 3절에서는 Kraus (2014)의 AUC 통계량을 이용한 연구 방법을 확장하여 다양하게 변형된 로짓 함수에 AUC 접근방법을 적용하고 이를 이용하여 얻은 모수 추정값과 기존의 로지스틱모형의 모수 추정값의 결과를 분석한다. 마지막 4절 결론에서 본 연구에서 제안된 연결함수에 AUC 접근방법을 이용하여 얻은 결과에 대하여 정리하고 결론을 유도한다.

2. 불균형 자료와 변형된 로짓 연결함수

Brown과 Mues (2012)는 신용평가 관점에서 불균형 자료를 부도의 관측값이 정상의 관측값보다 매우 작게 관측된 자료라고 정의하였고, 실제로 관측된 자료에 대하여 로지스틱 회귀모형, 의사결정나무, 뉴럴 네트워크(Neual networks), 선형판별분석(Linear discriminant analysis) 등의 일곱 가지 방법론의 판별력을 비교하였다. 부도의 비율이 30%인 경우에 로지스틱 모형으로 얻어진 AUC가 다른 방법으로 얻어진 AUC와 비슷한 결과를 보였으나, 부도의 비율이 15%, 10%, 5% 등으로 낮아질수록 로지스틱 모형의 AUC는 다른 방법의 AUC보다 낮아 분류 성능도 낮아지는 것을 확인하였다. Calabrese와 Osmetti (2011)은 로지스틱 모형의 추정값의 적게 관측된 종속변수의 발생 확률을 과소 추정하는 문제점과 좌우가 대칭인 로짓 연결함수는 비대칭 형태의 반응곡선을 제대로 적합하지 못하는 점을 언급하였다. Allison (2008)은 불균형 자료에 대해 SAS를 이용한 로지스틱 모형의 모수 추정의 경우를 설명하였을 때 MLE의 수렴 실패의 문제를 지적하였다. 이러한 MLE의 수렴 실패는 모수 추정의 안정성에 문제가 있으며, 이러한 문제는 자료의 표본의 수가 작은 경우 또는 반응변수의 빈도의 분포가 명확하게 나누어지는 경우에 흔히 발생하는 것으로 알려졌다.

불균형 자료의 모수 추정에 대한 로지스틱 방법의 문제점을 보완하기 위하여 Pepe 등 (2005)이 제안한 AUC 접근방법을 이용한다. 낮은 부도율 하에서 두 방법을 비교하기 위하여 다음과 같은 변형된 로짓



(a) Probability levels for fixed $b = 1$ (b) Probability levels for fixed $a = 1$

Figure 2.1. Probability levels with varying a and b.

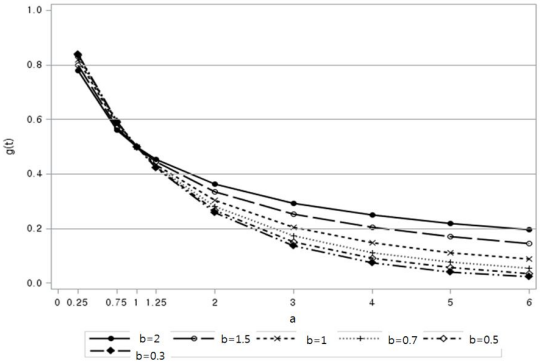


Figure 2.2. Probability levels with simultaneously varying a and b.

연결함수를 고려한다.

$$g(t) = ((1 + \exp(-bt))^a)^{-1}, \quad a > 0,\ b > 0. \tag{2.1}$$

Burr (1942)는 비대칭적인 로짓 분포의 모수를 추정하기 위하여 a값의 변화에 따라 반응 곡선의 첨도가 변화하는 분포 $F(z; a) = ((1 + \exp(-z))^a)^{-1}$ 제안하였다. 식 (2.1)에서 $t = X^T\beta$라 하면, 이것은 식 (1.1)의 우변과 같고 이는 부도율을 의미한다.

Figure 2.1의 (a)는 X가 다변량 표준정규분포로부터 생성된 설명변수이고 $t = X^T\beta$일 때, 식 (2.1)의 고정된 $b = 1$에 대하여 a가 변화하였을 때 $g(t)$값을 나타낸 것으로 $a = 1$인 경우는 로짓 연결함수인 경우이며, a값이 커질수록 $t \geq 0$에서 빠르게 1로 수렴하고 $t < 0$에서 느리게 0으로 수렴하는 경향을 보인다. Figure 2.1의 (b)는 식 (2.1)의 고정된 $a = 1$에 대하여 b가 변화하였을 때 $g(t)$의 값으로, $b = 1$인 경우는 로짓 연결함수의 경우이며 b값이 커질수록 $t \geq 0$과 $t < 0$ 에서 각각 1과 0으로 빠르게 수렴하는 결과를 보인다. Figure 2.2는 식 (2.1)의 a와 b가 동시에 변화할 때의 $g(t)$값을 나타낸 것으로 $a \geq 1$에 대하여 b가 작아질수록 $g(t)$의 값이 작아지는 경향을 보였고, $a < 1$에 대하여 b가 작아질수록 $g(t)$는 커지는 경향을 나타냈다. Table 2.1은 1,000,000개의 표본을 생성하여 a와 b값에 대응하는 $g(t)$의 값으로, 그 결과는 Figure 2.2와 동일한 경향을 보인다. 본 연구에서는 회색으로 표시된 셀을 참고하여 a와 b 값을 설정하여 모의실험에서 낮은 부도율의 상황을 가정한다.

Table 2.1. Obtained probability for a and b with 1,000,000 samples

a	$b=2$	$b=1.5$	$b=1$	$b=0.7$	$b=0.5$	$b=0.3$
0.25	0.7801	0.8000	0.8193	0.8293	0.8346	0.8385
0.75	0.5617	0.5707	0.5809	0.5868	0.5903	0.5929
1	0.4997	0.4998	0.4998	0.4998	0.4999	0.4999
1.25	0.4530	0.4449	0.4351	0.4289	0.4252	0.4223
2	0.3618	0.3357	0.3030	0.2815	0.2680	0.2571
3	0.2928	0.2537	0.2046	0.1723	0.1520	0.1356
4	0.2497	0.2039	0.1479	0.1121	0.0904	0.0733
5	0.2194	0.1702	0.1119	0.0765	0.0559	0.0405
6	0.1967	0.1458	0.0876	0.0541	0.0357	0.0228

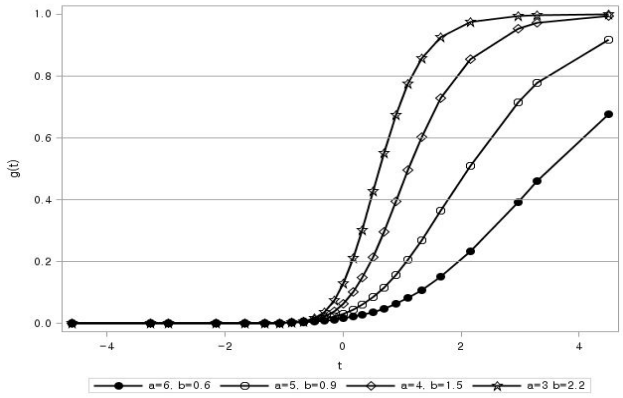


Figure 3.1. Logit functions with varying a and b.

3. 모의실험

AUC 접근방법으로 얻어진 추정량과 로지스틱모형을 이용한 최대가능도 추정량에 대한 비교를 모의실험한 Kraus (2014)의 방법을 확장하여, 식 (2.1)을 연결함수로 이용한 낮은 부도율 상황의 불균형 데이터를 생성하고 이에 대하여 두 방법으로 얻어진 모수 추정량을 비교한다. 설명변수 $X_1,\ X_2,\ X_3$를 평균이 0이고 분산이 1인 다변량 표준정규분포에서 각각 $n_1=500,\ n_2=300,\ n_3=200$개의 자료를 생성한다. 초기 모수 $\beta=(1,0.5,0.3)$와 설명변수 X의 선형결합 $t=X^T\beta$를 이용하여 식 (2.1)의 $g(t)$를 연결함수로 이용하여 이항 종속변수를 생성한다. 이 때 식 (2.1)의 a와 b를 다음과 같은 네 가지 경우로 고려하여 초기 모수추정에 대한 두 방법의 추정량의 결과를 비교한다.

Case 1) $p_1=0.05$ $(a=6,\ b=0.65)$, Case 3) $p_3=0.20$ $(a=4,\ b=1.5)$,

Case 2) $p_2=0.10$ $(a=5,\ b=0.9)$, Case 4) $p_4=0.30$ $(a=3,\ b=2.2)$.

Figure 3.1은 각 경우에 대응하는 p_i는 $g(t)$의 평균값을 나타낸 것으로 모든 경우가 비대칭적인 곡선의 형태를 나타내고, a가 크고 b가 작을수록 $g(t)$값도 작아지고 $a=6,\ b=0.65$이면 부도의 확률은 약 0.05이고 정상과 부도가 95 : 5인 비율을 나타낸다. $g(t)=\pi$로 설정하고 이항분포 Ber(π)로 부터 종속변수 $Y=\{1,0\}$을 생성한다. 변형된 로짓 함수의 $a,\ b$값에 따른 $p_i,\ i=1,2,3,4$별 로지스틱모형을 이용한 최대가능도 추정방법(ML 방법)의 모수 추정 결과와 AUC 접근방법의 모수 추정 결과를 정리하였다.

Table 3.1. Coefficients and AUC for ML method and AUC approach ($n = 500$)

$n = 500$	$\widehat{\beta}_2^{ML}$	$\widehat{\beta}_2^{AUC}$	$\widehat{\beta}_3^{ML}$	$\widehat{\beta}_3^{AUC}$	$\widehat{\text{AUC}}^{ML}$	$\widehat{\text{AUC}}^{AUC}$	$\widehat{\text{AUC}}^{AUC} - \widehat{\text{AUC}}^{ML}$
Case 1)	0.5126	0.5096	0.3191	0.3041	0.8705	0.8730	0.0025
$p_1 = 0.05$	(0.1522)	(0.1773)	(0.1938)	(0.2140)	(0.0305)	(0.0299)	
Case 2)	0.5081	0.4994	0.2910	0.2948	0.8816	0.8830	0.0014
$p_2 = 0.10$	(0.1277)	(0.1367)	(0.1337)	(0.1506)	(0.0228)	(0.0226)	
Case 3)	0.4988	0.5021	0.3055	0.3024	0.9180	0.9185	0.0005
$p_3 = 0.20$	(0.0763)	(0.0774)	(0.0670)	(0.0696)	(0.0122)	(0.0122)	
Case 4)	0.4992	0.4996	0.3003	0.2997	0.9426	0.9249	0.0003
$p_4 = 0.30$	(0.0583)	(0.0612)	(0.0552)	(0.0565)	(0.0090)	(0.0090)	

AUC = area under the ROC curve, ML = maximum likelihood.

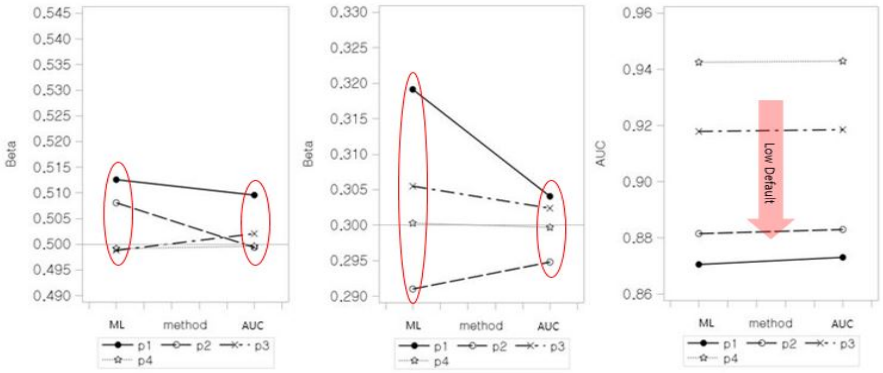


Figure 3.2. Coefficients and AUC for ML method and AUC approach ($n = 500$).

먼저 $n = 500$인 경우, Table 3.1에 각각 ML 방법과 AUC 접근방법의 100번의 반복에 대한 추정값의 평균을 구하고 표준편차를 괄호 안의 값으로 나타냈다. 먼저 부도의 확률이 $p_1 = 0.05$인 경우의 ML 방법의 모수 추정결과로 $\widehat{\beta}_2^{ML}$과 $\widehat{\beta}_3^{ML}$의 평균은 각각 0.5126과 0.3191이며, AUC 접근방법의 결과로 $\widehat{\beta}_2^{AUC}$과 $\widehat{\beta}_3^{AUC}$의 평균은 각각 0.5096과 0.3041로 두 방법 모두 원래의 모수 0.5와 0.3과는 약간의 편의를 보였다. 부도율 $p_2 = 0.10$에 대하여 ML 방법의 추정값의 평균은 각각 0.5081, 0.2910이고, AUC 접근방법의 경우 각각 0.4994와 0.2948로 나타났다. 또한 부도율이 $p_3 = 0.20$인 경우 ML 방법의 추정값의 평균은 0.4988과 0.3055였고, AUC 접근방법의 경우 각각 0.5021, 0.3024로 나타났다. $p_4 = 0.30$인 경우 ML 방법의 평균이 0.4992, 0.3003이고, AUC 접근방법의 평균이 0.4996, 0.2997이었다. 위의 결과로 특히 부도의 확률이 $p_1 = 0.05$와 $p_2 = 0.10$일 때 다른 두 부도율에 비하여 ML 방법보다 AUC 접근방법이 실제 모수를 보다 정확히 추정하는 것을 탐색할 수 있다($\widehat{\beta}$에 강조된 부분).

$p_1 = 0.05$인 경우 $\widehat{\beta}_2^{ML}$과 $\widehat{\beta}_3^{ML}$의 표준편차는 각각 0.1522, 0.1938로 $\widehat{\beta}_2^{AUC}$와 $\widehat{\beta}_3^{AUC}$의 표준편차 0.1773, 0.2140보다 작은 값을 보였다. 다른 세 부도율의 경우도 ML 방법의 추정값의 표준편차가 AUC 접근방법 추정값의 표준편차보다 작은 결과를 보였으므로 모두의 경우에서 AUC 접근방법의 모수 추정이 효율적인 결과를 보이지는 않았다. 분류성과 기준에서 AUC 접근방법으로 구한 AUC 추정값의 평균 $\widehat{\text{AUC}}^{AUC}$가 각각 0.8730, 0.8830, 0.9185, 0.9249로 ML 방법으로 구한 AUC 추정값의 평균 $\widehat{\text{AUC}}^{ML}$보다 각각의 p_i들에 대하여 모두 큰 값을 가졌다.

Figure 3.2는 위의 Table 3.1의 결과를 탐색적으로 표현한 것으로 왼쪽 그림은 모수 $\beta_2 = 0.5$가 참조

Table 3.2. Coefficients and AUC for ML method and AUC approach ($n = 300$)

$n = 300$	$\widehat{\beta}_2^{ML}$	$\widehat{\beta}_2^{AUC}$	$\widehat{\beta}_3^{ML}$	$\widehat{\beta}_3^{AUC}$	$\widehat{\text{AUC}}^{ML}$	$\widehat{\text{AUC}}^{AUC}$	$\widehat{\text{AUC}}^{AUC} - \widehat{\text{AUC}}^{ML}$
Case 1)	0.5416	0.5276	0.3293	0.3265	0.8764	0.8807	0.0043
$p_1 = 0.05$	(0.2326)	(0.2541)	(0.2042)	(0.2384)	(0.0407)	(0.0399)	
Case 2)	0.4931	0.5036	0.3047	0.2984	0.8830	0.8852	0.0022
$p_2 = 0.10$	(0.1539)	(0.1638)	(0.1538)	(0.1493)	(0.0281)	(0.0276)	
Case 3)	0.4925	0.5035	0.2911	0.2934	0.9162	0.9172	0.0010
$p_3 = 0.20$	(0.0986)	(0.1148)	(0.0801)	(0.0892)	(0.0175)	(0.0174)	
Case 4)	0.5121	0.5090	0.2945	0.2959	0.9399	0.9405	0.0006
$p_4 = 0.30$	(0.0791)	(0.0835)	(0.0623)	(0.0688)	(0.0114)	(0.0114)	

AUC = area under the ROC curve, ML = maximum likelihood.

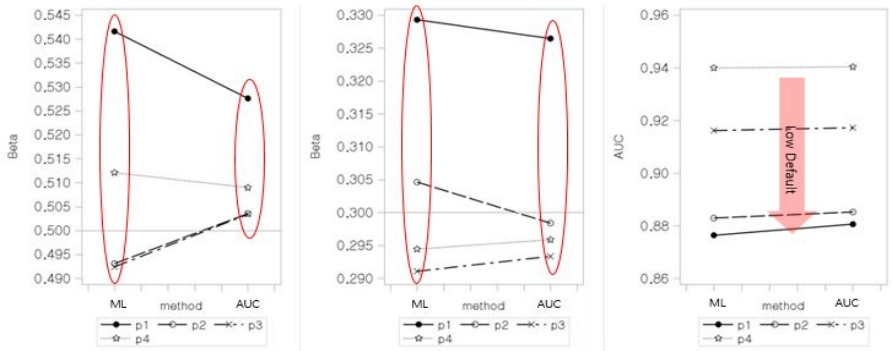


Figure 3.3. Coefficients and AUC for ML method and AUC approach ($n = 300$).

선인 실선으로 나타나 있고 부도율이 p_1, p_2인 경우에 ML 방법에 비해 AUC 접근방법의 추정값이 모수에 가깝게 추정된 것을 파악할 수 있다. 가운데 그림은 모수 $\beta_3 = 0.3$이 참조선으로 표현되어 있고 p_1, p_2경우에 AUC 접근방법의 추정값이 모수를 근접하게 추정하는 것으로 나타나 있다. 오른쪽 그림은 두 접근방법으로 AUC의 평균값을 나타낸 것으로 AUC 접근방법의 값이 모든 경우에 대하여 큰 것으로 나타났고, p_i가 낮아질수록 AUC도 낮아져 분류성과가 나빠지는 것으로 요약되어 있다.

$n = 300$인 경우에 대한 결과를 Table 3.2에 정리하였다. $p_1 = 0.05$일 때 두 방법 모두 편의가 가장 크게 나타났다. β_2의 경우 부도율이 p_1, p_3인 경우에 AUC 접근방법이 모수값 0.5에 보다 가깝게 추정하였고, β_3의 경우 p_2, p_3의 경우에 ML 방법보다 AUC 접근방법의 추정값이 0.3에 가깝게 추정하는 결과를 보였다. 추정값의 편차는 두 방법 모두 p_i가 낮아질수록 커지는 경향을 보여 낮은 부도율 p_1, p_2에서 p_3, p_4에 비하여 상대적으로 불안정한 추정결과를 나타냈다. 앞의 Table 3.1과 비교하여 n이 작아지면서 추정값의 편의가 커지는 경향을 확인할 수 있다. AUC 접근방법으로 계산된 AUC는 ML 방법으로 계산된 AUC보다 큰 것을 알 수 있고, Table 3.1과 비교해 두 AUC의 차이가 커진 것을 파악할 수 있다.

Figure 3.3의 β_2와 β_3의 추정값을 살펴보면, AUC 접근방법이 모수 0.5와 0.3에 ML 방법과 동등하거나 모수에 근접한 추정값을 나타내는 것을 파악할 수 있다. Figure 3.2와 비교하여 참조선을 기준으로 흩어진 정도가 심해져 편의가 커진 것을 탐색할 수 있다. Figure 3.3의 AUC 추정값을 살펴보면, 앞의 Figure 3.2와 비교하여 전체적으로 AUC의 평균값이 커지고 여전히 AUC 접근방법이 ML 방법보다 큰 값을 보였다.

Table 3.3. Coefficients and AUC for ML method and AUC approach ($n = 200$)

$n = 200$	$\widehat{\beta}_2^{ML}$	$\widehat{\beta}_2^{AUC}$	$\widehat{\beta}_3^{ML}$	$\widehat{\beta}_3^{AUC}$	$\widehat{\text{AUC}}^{ML}$	$\widehat{\text{AUC}}^{AUC}$	$\widehat{\text{AUC}}^{AUC} - \widehat{\text{AUC}}^{ML}$
Case 1)	0.5272	0.5172	0.3237	0.3182	0.8657	0.8721	0.0064
$p_1 = 0.05$	**(0.4555)**	**(0.3815)**	**(0.3761)**	**(0.3491)**	(0.0548)	(0.0534)	
Case 2)	0.5217	0.5056	0.3054	0.3010	0.8879	0.8914	0.0035
$p_1 = 0.10$	(0.2043)	(0.2043)	(0.1924)	(0.2165)	(0.0350)	(0.0340)	
Case 3)	0.5121	0.5117	0.3050	0.3035	0.9211	0.9227	0.0016
$p_3 = 0.20$	(0.1488)	(0.1625)	(0.0978)	(0.1070)	(0.0228)	(0.0227)	
Case 4)	0.5079	0.5023	0.3146	0.3120	0.9401	0.9410	0.0009
$p_4 = 0.30$	(0.0862)	(0.0912)	(0.0997)	(0.1040)	(0.0150)	(0.0149)	

AUC = area under the ROC curve, ML = maximum likelihood.

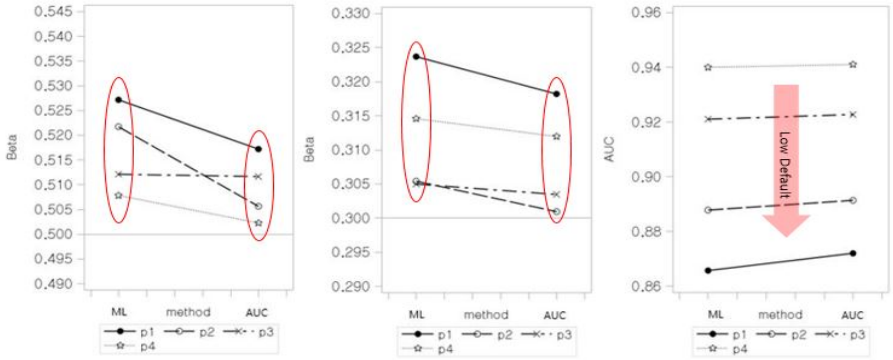


Figure 3.4. Coefficients and AUC for ML method and AUC approach ($n = 200$).

Table 3.3은 $n = 200$인 경우의 결과를 나타냈다. p_1에서 $\widehat{\beta}_2^{ML}$와 $\widehat{\beta}_3^{ML}$이 각각 0.5272, 0.3237이었고 $\widehat{\beta}_2^{AUC}$와 $\widehat{\beta}_3^{AUC}$는 각각 0.5172, 0.3182로 다른 경우에 비하여 편의는 가장 컸으나 AUC 접근방법이 ML 방법보다 뛰어난 추정 결과를 보였다. p_2의 경우 $\widehat{\beta}_2^{ML}$, $\widehat{\beta}_3^{ML}$은 각각 0.5217, 0.3054였고, $\widehat{\beta}_2^{AUC}$와 $\widehat{\beta}_3^{AUC}$가 각각 0.5056, 0.3010으로 p_3, p_4에 비하여 AUC 접근방법의 추정값이 모수를 잘 추정하는 것으로 나타났다. p_i가 낮을수록 추정값의 편차가 크게 나타났고, 앞의 Table 3.2와 비교하면 n이 작아지면서 추정값의 편차가 커져 불안정한 모수 추정 결과를 보였다. 그러나 앞의 결과와 달리 $p_1 = 0.05$에서 AUC 접근방법의 모수 추정값의 편차가 ML 방법보다 작게 나왔다. p_i가 커지면서 AUC도 커져 분류성과도 좋아졌다. 여전히 AUC 접근방법으로 계산된 AUC가 크게 나왔고 두 방법의 AUC의 차이가 앞의 Table 3.2와 비교하여 커진 것을 확인한다.

Figure 3.4를 살펴보면 두 방법 모두 0.5와 0.3 참조선에서 떨어져 있어 편의를 확인할 수 있지만, AUC 접근방법의 추정값이 대체로 참조선에 보다 가깝게 나타났다. p_i가 낮을수록 AUC 값도 낮은 값을 보이는 것이 나타났고, 앞의 두 Figure 3.2와 Figure 3.3과 비교하여 모수를 크게 추정하는 경향과 전반적인 AUC의 상승을 탐색할 수 있다.

4. 결론

단기간의 일반적인 평가시스템에 있어서 부도의 수가 희박하고 이러한 불균형 자료에 대한 로지스틱 모형의 판별력에 대한 문제점이 지적되었다 (Tasche, 2009; Calabrese와 Osmetti, 2011). Pepe 등

(2005)는 AUC 접근방법이 로지스틱 가정이 맞지 않는 상황에서 좋은 추정방법임을 제안하였고, Kraus (2014)는 일반적인 상황으로 확장하여 로짓 이외에 프로빗, complementary log-log 등의 연결함수를 고려하여 Pepe 등 (2005)의 AUC 접근방법이 로지스틱모형의 최대가능도 추정량보다 좋은 결과를 보이는 것을 발견하였다.

본 연구에서는 신용평가 측면에서 부도율이 낮은 상황을 가정하고 Pepe 등 (2005)의 AUC 접근방법을 이용하였다. 부도율이 낮을 때 얻어지는 불균형 자료를 생성하기 위해 변형된 로짓 연결함수를 이용하였다. 변형된 로짓 연결함수로 정상과 부도의 비율이 각각 95 : 5, 9 : 1, 8 : 2, 7 : 3인 불균형 자료를 생성하고 AUC 접근방법과 기존의 ML 방법과 비교한 결과를 얻었다.

정상과 부도의 비율이 8 : 2, 7 : 3인 경우에 AUC 접근방법은 ML 방법과 유사한 추정 결과를 보였다. 부도율이 보다 낮은 경우에 대하여 두 방법은 모두 추정값의 편의를 보였고 편차도 커지는 경향이 있었으나, AUC 접근방법이 ML 방법 보다는 모수에 가까운 추정결과를 보였다. 동일한 부도의 비율로 관측값의 수를 줄였을 때, 두 방법 모두 추정값의 편의와 편차가 커져 불안정한 추정 경향을 보였다. 그럼에도 여전히 AUC 접근방법이 ML 방법보다 비교적 편의가 작고 AUC가 큰 값을 보였으므로 적은 수의 자료에 대하여 좋은 결과를 보였다. 즉 적은 관측 수의 불균형 자료에 대해 Pepe 등 (2005)의 AUC 접근방법이 ML 방법보다 더 나은 모수 추정 결과를 보이는 것을 확인하였다.

References

Allison, P. D. (2008). Convergence failures in logistic regression, *In SAS Global Forum*, **360**, 1–11.

Bamber, D. C. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, *Journal of Mathematical Psychology*, **12**, 387–415.

Brown, I. and Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets, *Expert Systems with Applications*, **39**, 3446–3453.

Burr, I. W. (1942). Cumulative frequency functions, *The Annals of Mathematical Statistics*, **13**, 215–232.

Calabrese, R. and Osmetti, S. A. (2011). Generalized extreme value regression for binary rare events data: an application to credit defaults, *Bulletin of the International Statistical Institute LXII*, 58th Session of the International Statistical Institute, 5631–5634.

Cavanagh, C. and Sherman, R. P. (1998). Rank estimators for monotonic index models, *Journal of Econometrics*, **84**, 351–381.

Dreiseitl, S., Ohno-Machado, L., and Binder, M. (2000). Comparing three-class diagnostic tests by three-way ROC analysis, *Medical Decision Making*, **20**, 323–331.

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

Engelmann, B., Hayden, E., and Tasche, D. (2003). Measuring the discriminative power of rating systems, *Risk*, 82–86.

Fawcett, T. (2003). ROC graphs: Notes and practical considerations for data mining researchers, HP Labs Technical Report HPL-2003-4, CA, USA.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model, the maximum rank correlation estimator, *Journal of Economics*, **35**, 303–316.

Heckerling, P. S. (2001). Parametric three-way receiver operating characteristic surface analysis using mathematica, *Medical Decision Making*, **21**, 409–417.

Hong, C. S. and Cho, M. H. (2015a). VUS and HUM represented with Mann-Whitney statistic, *Communications for Statistical Applications and Methods*, **22**, 223–232.

Hong, C. S. and Cho, M. H. (2015b). Test statistics for volume under the ROC surface and hypervolume under the ROC manifold, *Communications for Statistical Applications and Methods*, **22**, 377–387.

Hong, C. S. and Choi, J. S. (2009). Optimal threshold from ROC and CAP curves, *The Korean Journal of Applied Statistics*, **22**, 911–921.

Hong, C. S., Joo, J. S., and Choi, J. S. (2010). Optimal thresholds from mixture distributions, *The Korean*

Journal of Applied Statistics, **23**, 13–28.

Hong, C. S. and Jung, D. G. (2014). Standard criterion of hypervolume under the ROC manifold, *Journal of the Korean Data & Information Science Society*, **25**, 473–483.

Hong, C. S. and Jung, E. S. (2013). Optimal thresholds criteria for ROC surfaces, *Journal of The Korean Data and Information Science Society*, **24**, 1489–1496.

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**, 1–8.

Hong, C. S., Won, C. H., and Jeong, D. G. (2015). Parameter estimation of linear function using VUS and HUM maximization, *Journal of the Korean Data & Information Science Society*, To appear.

Hong, C. S. and Wu, Zhi Qiang (2014). Alternative accuracy for multiple ROC analysis, *Journal of The Korean Data & Information Science Society*, **25**, 1521-1530.

Hosmer, D. W. (2000). *Applied Logistic Regression*, 2nd ed., Wiley, New York.

Joseph, M. P. (2005). A PD validation framework for Basel II internal ratings-based systems, *Quantitative Analyst Basel II Project*, Commonwealth Bank of Australia.

Kraus, A. (2014). *Recent Methods from Statistics and Machine Learning for Credit Scoring*, Dissertation an der Fakultät für Mathematik, Informatik und Statistik, der Ludwig-Maximilians-Universität München, München; http://edoc.ub.uni-muenchen.de/17143/1/Kraus_Anne.pdf.

Li, J. and Fine, J. P. (2008). ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies, *Biostatistics*, **9**, 566–576.

Mossman, D. (1999). Three-way ROCs, *Medical Decision Making*, **19**, 78–89.

Nakas, C. T., Alonzo, T. A., and Yiannoutsos, C. T. (2010). Accuracy and cut off point selection in three class classification problems using a generalization of the Youden index, *Statistics in Medicine*, **29**, 2946–2955.

Nakas, C. T. and Yiannoutsos, C. T. (2004). Ordered multiple-class ROC analysis with continuous measurements, *Statistics in Medicine*, **23**, 3437–3449.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization, *The Computer Journal*, **7**, 308–313.

Patel, A. C. and Markey, M. K. (2005). Comparison of three-class classification performance metrics: A case study in breast cancer CAD, *International Society for Optical Engineering*, **5749**, 581–589.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford UniversityPress, Oxford.

Pepe, M. S., Cai, T., and Longton, G. (2005). Combining predictors for classification using the area under the receiver operating characteristic curve, *Biometrics*, **1**, 221–229.

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

Scurfield, B. K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability, *Journal of Mathematical Psychology*, **40**, 253–269.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator, *Econometrics*, **61**, 123–137.

Sobehart, J. R. and Keenan, S. C. (2001). Measuring default accurately, Credit risk special report, *Risk*, **14**, 31–33.

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

Swets, J. A., Dawes, R. M., and Monahan, J. (2000). Better decisions through science, *Scientific American*, **283**, 82–87.

Tasche, D. (2009). Estimating discriminatory power and PD curves when the number of defaults is small, *Lioyds Banking Group*.

Wandishin, M. S. and Mullen, S. J. (2009). Multiclass ROC analysis, *Weather and Forecasting*, **24**, 530–547.

Zou, K. H., O'Malley, A. J., and Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models, *Circulation*, **115**, 654–657.

AUC 최적화를 이용한 낮은 부도율 자료의 모수추정

홍종선[a,1] · 원치환[a]

[a]성균관대학교 통계학과

(2015년 11월 2일 접수, 2015년 12월 28일 수정, 2016년 1월 5일 채택)

요 약

이항 분류모형에서 선형 스코어의 함수인 리스크 스코어를 고려하고, 선형 스코어의 계수를 추정하는 문제를 고려한다. 계수를 추정하는 대표적인 방법으로 로지스틱모형을 이용하는 방법과 AUC를 최대화하여 구하는 방법이 있다. AUC 접근방법으로 구한 모수 추정량은 로지스틱모형을 이용한 선형 스코어의 모수의 최대가능도 추정량보다 자료가 로지스틱 가정이 맞지 않는 일반적인 상황에서도 좋은 추정 결과를 보인다. 본 연구에서는 신용평가모형에서 흔히 접하는 정상보다 부도 경우가 현저하게 작은 상태인 낮은 부도율의 자료를 고려하고, 낮은 부도율의 자료에 AUC 접근방법을 적용한다. 부도의 비율이 정상의 비율보다 현저하게 낮은 불균형 자료를 생성하기 위하여 수정된 로짓 함수를 연결함수로 사용한다. 낮은 부도율의 상황인 불균형 자료에 AUC 접근방법을 적용한 판별결과가 로지스틱 모형 추정방법보다 동등하거나 더 나은 모수추정 결과를 보이는 것을 확인하였다.

주요용0어: 분류점, ROC 곡선, 연결함수, 위험, 판별

[1]교신저자: (110-745) 서울 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: cshong@skku.edu