Journal of the Korean Data & Information Science Society
2015, **26**(6), 1305–1315

http://dx.doi.org/10.7465/jkdi.2015.26.6.1305
한국데이터정보과학회지

VUS와 HUM 최적화를 이용한 선형함수의 모수추정

홍종선[1] · 원치환[2] · 정동길[3]

[123]성균관대학교 통계학과

접수 2015년 8월 3일, 수정 2015년 9월 8일, 게재확정 2015년 9월 14일

요 약

ROC 곡선을 구성하는 한 개의 스코어 변수로 이루어진 분류모형을 확장하여 선형 스코어의 함수인 리스크 스코어를 고려하고, 선형 스코어의 계수를 추정하기 위한 방법으로 AUC를 최대화하는 방법을 사용한다. 이런 AUC 접근방법으로 구한 스코어의 계수 추정량은 로지스틱모형을 이용한 선형 스코어의 모수의 최대가능도 추정량보다 자료가 로지스틱 가정이 맞지 않는 일반적인 상황에서도 좋은 추정 결과를 보인다. 본 연구에서는 다항범주로 분류되어 현실적인 판별 및 예측 상황을 고려하여 AUC 접근방법을 확장한 VUS와 HUM 접근방법을 제안한다. 연결함수로는 로짓, complementary log-log와 로짓을 변형한 함수의 세 종류와 그리고 다양한 분류점의 분포인 경우에 대하여도 모의실험을 실시하였다. 본 논문에서는 다항범주 판별결과에 대하여 VUS와 HUM 접근방법도 AUC 접근방법과 유사하게 다양한 연결함수에 대하여 로지스틱모형 추정방법보다 동등하거나 더 나은 모수추정 결과를 보이는 것을 확인하였다.

주요용어: 곡면, 다면체, 분류점, 스코어, 연결함수, 위험, 판별.

1. 서론

ROC (receiver operating characteristic) 곡선은 의학진단이나 신용평가에서 모형의 성능 (performance)을 탐색하는 유용한 시각적인 방법이다. ROC 곡선은 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.

의학진단 또는 신용평가 측면에서 분류하기 위한 확률변수 X를 스코어 변수라 하면 분류점 c로부터의 이항 결과를 다음과 같이 $Y = \{0, 1\}$으로 나타낼 수 있다. 이 때의 Y의 원소 0은 환자나 차주의 정상상태 (good, non-default), 1은 부도상태 (bad, default)로 정의할 수 있으며 민감도 (sensitivity)와 1-특이도 (1-specificity)는 각각 다음과 같은 확률로 정의한다.

$$sens(c) = P(X \geq c|Y = 1),$$
$$1 - spec(c) = P(X \geq c|Y = 0).$$

[1] 교신저자: (110-745) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과, 교수.
E-mail: cshong@skku.ac.kr
[2] (110-745) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과, 대학원생.
[3] (110-745) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과, 대학원생.

스코어 변수에 대한 ROC 곡선은 가능한 모든 분류점 c에 대한 민감도와 1-특이도의 집합으로 다음과 같이 표현된다.

$$ROC(c) = \{(1 - spec(c), sens(c)),\ c \in (-\infty, \infty)\}. \tag{1.1}$$

ROC 곡선의 평가기준 측도로 ROC 곡선의 아래의 면적을 계산한 AUC (the 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 등 (2005)은 기존의 한 개의 스코어 변수로 이루어진 분류모형에서 여러 개의 변수들의 선형 결합으로 이루어진 스코어 함수로 확대하여 고려하였다. 선형 스코어 (linear score)를 다음과 같이 정의하였고, 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 곡선을 식 (1.1)과 동일하게 정의하였다.

$$sens(c) = P(L_\beta(X_i^1) \geq c),$$
$$1 - spec(c) = P(L_\beta(X_j^2) \geq c).$$

Pepe (2003)는 리스크 스코어 (risk score)를 선형 스코어의 함수로 다음과 같이 정의하고,

$$P(Y = 1|X) = g(L_\beta(X)), \tag{1.2}$$

함수 $g(\cdot)$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. 그리고 AUC 통계량을 목적함수로 설정하여 AUC를 최대화하는 β를 추정하였다. 이 때의 목적함수인 경험적 AUC 통계량은 식 (1.3)과 같고, 이것은 Mann-Whitney 통계량과 동등하다 (Bamber, 1975).

$$\widehat{AUC}(b) = \frac{1}{n^1 n^2} \sum_{i=1}^{n^1} \sum_{j=1}^{n^2} \left\{ I[L_b(X_i^1) > L_b(X_j^2)] + 0.5 I[L_b(X_i^1) = L_b(X_j^2)] \right\}, \tag{1.3}$$

그리고 AUC 통계량을 목적함수로 하여 모수 β를 추정하는 방법은 식 (1.4)와 같고 이 방법은 AUC 접근방법이라고 한다.

$$\widehat{\beta}^{AUC} = argmax \widehat{AUC}(b). \tag{1.4}$$

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 가정이 맞지 않는 일반적인 상황에서도 여전히 로지스틱모형의 추정량보다 동등하거나 더 좋은 것이라고 주장하였다. 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.

본 연구에서는 두 가지 범주로 분류되어 있는 상황보다 현실적인 상황인 세 가지 이상의 범주로 분류되어 있는 상황을 고려한다. 일반적인 신용평가 관점에서 ROC 곡선은 두 가지 범주인 부도 혹은 정상을 분류하는 모형의 판별력을 나타내지만 세 가지 범주인 정상, 위험, 부도에 대하여 혹은 그 이상의 범주에 대하여 모형의 판별력을 측정하는것이 선호되므로 세 가지 이상의 범주에서 선형 모형의 판별력을 최대화시키는 모수추정에 관한 연구를 한다.

본 논문 2절에서는 분류결과가 3 이상인 분류자 모형을 가정하고, 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)를 설명하고 모수를 추정하기 위한 목적함수로 VUS와 HUM 통계량을 설정한다. 3절에서는 Kraus (2014)의 AUC 통계량을 이용한 연구 방법을 확장하여 VUS와 HUM 통계량을 이용한 VUS와 HUM 접근방법을 제안하고 이를 이용하여 얻은 모수 추정값과 기존의 로지스틱모형의 모수 추정값의 결과를 분석한다. 마지막 4절에서 본 연구에서 제안한 접근방법을 이용하여 얻은 결과에 대하여 정리하고 결론을 유도하고 향후 연구과제에 대해 토론한다.

2. VUS와 HUM을 이용한 모수추정

모형에서 결과를 두 가지 범주로 판별하는 경우에 ROC 곡선과 AUC 통계량이 사용되는 반면 모형에서 결과를 세 가지 범주로 판별하는 경우에 ROC 곡면에 대한 VUS 통계량과 네 가지 이상의 분류결과를 가지는 경우에 대하여 각각 ROC 다면체에 대한 HUM 통계량을 정의하였고, 이에 대한 판별력의 기준들이 제안되었다 (Scurfield, 1996; Mossmann, 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, 2014; Hong과 Cho, 2015).

분류결과가 3 이상인 경우 $Y = \{1, 2, 3, 4\}$에 대하여 설명변수 $X_i^1,\ X_j^2,\ X_k^3,\ X_l^4$는 각각 $i = 1, \ldots, n^1,\ j = 1, \ldots, n^2,\ k = 1, \ldots, n^3,\ l = 1, \ldots, n^4$개의 자료를 생성하고, 반응변수의 값 $Y = 1,\ Y = 2,\ Y = 3,\ Y = 4$로 설정하여 식 (1.2)을 확장한 식 (2.1)과 같은 분류자 모형을 표현한다.

$$\begin{aligned} P(Y_i = 1|X_i) &= g(X_{1i}^1 + \beta_2 X_{2i}^1 + \cdots + \beta_p X_{pi}^1) &&= g(L_\beta(X_i^1)), \\ P(Y_i = 2|X_j) &= g(X_{1j}^2 + \beta_2 X_{2j}^2 + \cdots + \beta_p X_{pj}^2) &&= g(L_\beta(X_j^2)), \\ P(Y_i = 3|X_k) &= g(X_{1k}^3 + \beta_2 X_{2k}^3 + \cdots + \beta_p X_{pk}^3) &&= g(L_\beta(X_k^3)), \\ P(Y_i = 4|X_l) &= g(X_{1l}^4 + \beta_2 X_{2l}^4 + \cdots + \beta_p X_{pl}^4) &&= g(L_\beta(X_l^4)). \end{aligned} \tag{2.1}$$

AUC 통계량에서 차원을 확장하여 VUS와 HUM 통계량을 다음과 같이 정의한다.

$$\begin{aligned} VUS(\beta) &= P(L_\beta(X_i^1) \geq L_\beta(X_j^2) \geq L_\beta(X_k^3)), \\ HUM(\beta) &= P(L_\beta(X_i^1) \geq L_\beta(X_j^2) \geq L_\beta(X_k^3) \geq L_\beta(X_l^4)). \end{aligned}$$

VUS와 HUM으로 정의된 목적함수를 식 (1.6)을 확장하여 각각 식 (2.2)와 식 (2.3)으로 설정한다.

$$\begin{aligned}\widehat{VUS}(b) = \frac{1}{n^1n^2n^3}\sum_{i=1}^{n^1}\sum_{j=1}^{n^2}\sum_{k=1}^{n^3}\{&I[L_b(X_i^1) > L_b(X_j^2) > L_b(X_k^3)] \\ &+ \frac{1}{2}I[L_b(X_i^1) = L_b(X_j^2) > L_b(_k^3)] \\ &+ \frac{1}{2}I[L_b(X_i^1) > L_b(X_j^2) = L_b(X_k^3)] \\ &+ \frac{1}{4}I[L_b(X_i^1) = L_b(X_j^2) = L_b(X_k^3)]\},\end{aligned} \tag{2.2}$$

$$\begin{aligned}\widehat{HUM}(b) = \frac{1}{n^1n^2n^3n^4}\sum_{i=1}^{n^1}\sum_{j=1}^{n^2}\sum_{k=1}^{n^3}\sum_{l=1}^{n^4}\{&I[L_b(X_i^1) > L_b(X_j^2) > L_b(X_k^3) > L_b(X_l^4)] \\ &+ \frac{1}{2}I[L_b(X_i^1) = L_b(X_j^2) > L_b(X_k^3) > L_b(X_l^4)] \\ &+ \frac{1}{2}I[L_b(X_i^1) > L_b(X_j^2) = L_b(X_k^3) > L_b(X_l^4)] \\ &+ \frac{1}{2}I[L_b(X_i^1) > L_b(X_j^2) > L_b(X_k^3) = L_b(X_l^4)] \\ &+ \frac{1}{4}I[L_b(X_i^1) = L_b(X_j^2) > L_b(X_k^3) = L_b(X_l^4)] \\ &+ \frac{1}{8}I[L_b(X_i^1) = L_b(X_j^2) = L_b(X_k^3) > L_b(X_l^4)] \\ &+ \frac{1}{8}I[L_b(X_i^1) > L_b(X_j^2) = L_b(X_k^3) = L_b(X_l^4)] \\ &+ \frac{1}{16}I[L_b(X_i^1) = L_b(X_j^2) = L_b(X_k^3) = L_b(X_l^4)]\}.\end{aligned} \tag{2.3}$$

VUS와 HUM이 최대가 되었을 때의 모수를 추정한다. 이와 같은 식 (2.4)와 식 (2.5)의 방법을 각각 VUS와 HUM 접근방법이라 한다.

$$\widehat{\beta}^{VUS} = argmax\widehat{VUS}(b), \tag{2.4}$$

$$\widehat{\beta}^{HUM} = argmax\widehat{HUM}(b). \tag{2.5}$$

본 연구에서는 Pepe (2003)와 Pepe 등 (2005)의 AUC 접근방법을 VUS와 HUM 접근방법으로 확장하여 모수를 추정한다. 목적함수 식 (2.2)와 식 (2.3)이 이산형이므로 Cavanagh와 Sherman (1998)이 적용한 NM알고리즘 (Nelder와 Mead, 1965)을 사용한다.

3. 모의실험

VUS와 HUM 접근방법에 대해 효율성을 살펴보기 위하여 우선 AUC 접근방법으로 얻어진 추정량과 로지스틱모형을 이용한 최대가능도 추정량에 대한 비교를 모의실험한 Kraus (2014)의 방법을 확장한다. 설명변수는 다음과 같은 분포에서 각각 100개의 자료를 생성한다.

$$X_1, X_2, X_3 \sim MVN\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}\right).$$

초기 모수 $\beta = (1, 0.5, 0.3)$으로 설정하여 $t = X^T\beta$를 생성하고, 식 (3.1)과 같은 세 종류의 연결함수를 이용하여 $g(t)$를 생성한다.

$$\begin{aligned} g_1(t) &= \frac{1}{1+exp(-t)}, && -\infty < t < \infty, \\ g_2(t) &= 1 - exp(-exp(t)), && -\infty < t < \infty, \\ g_3(t) &= \begin{cases} 1/(1+exp(-t)), & t \le 0, \\ 1/(1+exp(-8t)), & t > 0. \end{cases} \end{aligned} \tag{3.1}$$

Y의 범주가 0과 1 두 가지 일 때, 특수형태의 연결함수인 $g_3(t)$는 $t \le 0$에 대해 천천히 0으로 수렴하고, $t > 0$일 때는 빠르게 1로 수렴하는 형태로 0과 1로 수렴하는 속도가 다르며 특히 부도율이 낮을 때의 연결함수의 형태라고 볼 수 있다.

먼저 VUS를 이용한 접근방법으로 두 분류점 c_1과 c_2를 각각 $N(0.3, 0.1^2)$과 $N(0.7, 0.1^2)$에서 생성하고, $g(t) < c_1$이면 $Y = 1$, $c_1 \le g(t) < c_2$이면 $Y = 2$, 그리고 $c_2 \le g(t)$이면 $Y = 3$으로 설정한다. 각 연결함수별 로지스틱모형을 이용한 최대가능도 추정방법을 ML 방법이라 하고, ML 방법의 모수 추정 결과와 VUS 접근방법의 모수 추정 결과를 Table 3.1에 정리하였다.

Table 3.1 Coefficients and VUS for ML method and VUS approach ($\sigma = 0.1$)

Link	$\widehat{\beta}_2^{ML}$	$\widehat{\beta}_3^{ML}$	$\widehat{\beta}_2^{VUS}$	$\widehat{\beta}_3^{VUS}$	$\widehat{VUS}^{ML}$	$\widehat{VUS}^{VUS}$
$g_1(t)$	0.5226(0.1159)	0.3064(0.1040)	0.5278(0.1024)	0.3021(0.0822)	0.8779	0.8847
$g_2(t)$	0.5140(0.1384)	0.3177(0.1317)	0.5065(0.0808)	0.3033(0.0763)	0.9020	0.9160
$g_3(t)$	0.5205(0.1331)	0.3133(0.1006)	0.5059(0.0824)	0.2987(0.0737)	0.8891	0.9016

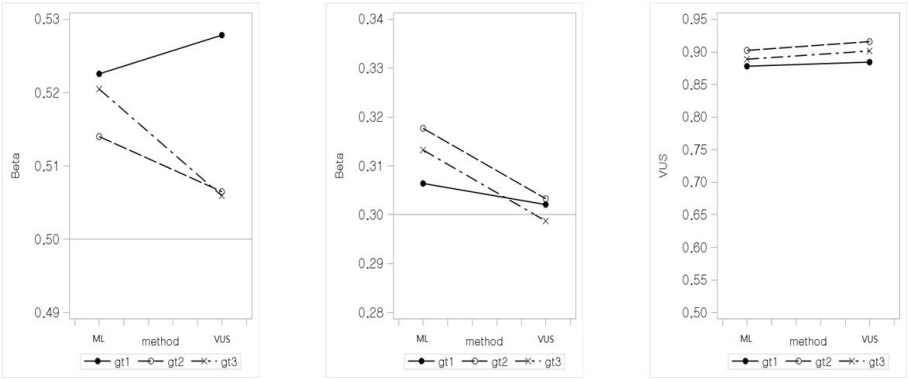

Figure 3.1 Coefficients and VUS for ML method and VUS approach ($\sigma = 0.1$)

Table 3.1에 각각 ML 방법과 VUS 접근방법의 1,000번의 반복에 대한 추정값의 평균을 구하고 표준편차를 괄호 안의 값으로 나타냈다. 먼저 연결함수가 $g_1(t)$인 경우의 ML 방법의 모수추정결과로 $\widehat{\beta}_2^{ML}$과 $\widehat{\beta}_3^{ML}$의 평균은 각각 0.5226과 0.3064이며, VUS 접근방법의 결과로 $\widehat{\beta}_2^{VUS}$과 $\widehat{\beta}_3^{VUS}$의 평균은 각각 0.5278과 0.3021으로 두 방법 모두 원래의 모수 0.5와 0.3에 가깝게 추정하는 것을 알 수 있다. 연결함수 $g_2(t)$에 대하여 ML 방법의 추정값의 평균은 각각 0.5140, 0.3177이고, VUS 접근방법의 경우 각각 0.5065와 0.3033으로 나타났다. 또한 연결함수가 $g_3(t)$인 경우 ML 방법의 추정값의 평균은 0.5205와 0.3130이었고, VUS 접근방법의 경우 각각 0.5059, 0.2987로 나타났다. 위의 결과로 특히 연

결함수가 $g_2(t)$와 $g_3(t)$일 때, ML 방법은 약간의 편의 (bias)를 보였고 VUS 접근방법은 실제 모수를 보다 정확히 추정하는 것을 알 수 있다.

연결함수가 $g_1(t)$인 경우 $\widehat{\beta}_2^{ML}$과 $\widehat{\beta}_3^{ML}$의 표준편차는 각각 0.1159, 0.1040으로 $\widehat{\beta}_2^{VUS}$과 $\widehat{\beta}_3^{VUS}$의 표준편차 0.1024, 0.0822보다 큰 값을 보였다. 다른 두 연결 함수의 경우도 ML 방법의 추정값의 표준편차가 VUS 접근방법 추정값의 표준편차보다 큰 결과를 보였으므로 세 연결함수 모두의 경우에서 VUS 접근방법의 모수추정이 효율적인 결과를 보이는 것을 파악할 수 있다.

분류성과 (performance)기준에서 VUS 접근방법으로 구한 VUS 추정값의 평균 $\widehat{VUS}^{VUS}$가 각각 0.8847, 0.9160, 0.9016으로 ML 방법으로 구한 VUS 추정값의 평균 $\widehat{VUS}^{ML}$보다 각각의 연결함수에 대하여 모두 큰 값을 가졌다.

Figure 3.1은 위의 Table 3.1의 결과를 탐색적으로 표현한 것으로 왼쪽 그림은 모수 $\beta_2 = 0.5$가 참조선인 실선으로 나타나 있고 연결함수가 $g_2(t)$, $g_3(t)$인 경우에 ML 방법에 비해 VUS 접근방법의 추정값이 모수에 가깝게 추정된 것을 파악할 수 있다. 가운데 그림은 모수 $\beta_3 = 0.3$가 참조선으로 표현되어 있고 세 연결함수 모두의 경우에 VUS 접근방법의 추정값이 모수를 근접하게 추정하는 것으로 나타나 있다. 오른쪽 그림은 두 접근방법으로 VUS의 평균값을 나타낸 것으로 VUS 접근방법의 값이 큰 것으로 요약되어 있다.

Table 3.2 Coefficients and VUS for ML method and VUS approach ($\sigma = 0.2$)

Link	$\widehat{\beta}_2^{ML}$	$\widehat{\beta}_3^{ML}$	$\widehat{\beta}_2^{VUS}$	$\widehat{\beta}_3^{VUS}$	$\widehat{VUS}^{ML}$	$\widehat{VUS}^{VUS}$
$g_1(t)$	0.5193(0.1627)	0.3316(0.1481)	0.5097(0.1510)	0.3335(0.1557)	0.6601	0.6693
$g_2(t)$	0.5245(0.1672)	0.3119(0.1542)	0.5236(0.1408)	0.3002(0.1185)	0.7256	0.7390
$g_3(t)$	0.5249(0.1418)	0.3138(0.1522)	0.5199(0.1046)	0.3089(0.1256)	0.7242	0.7389

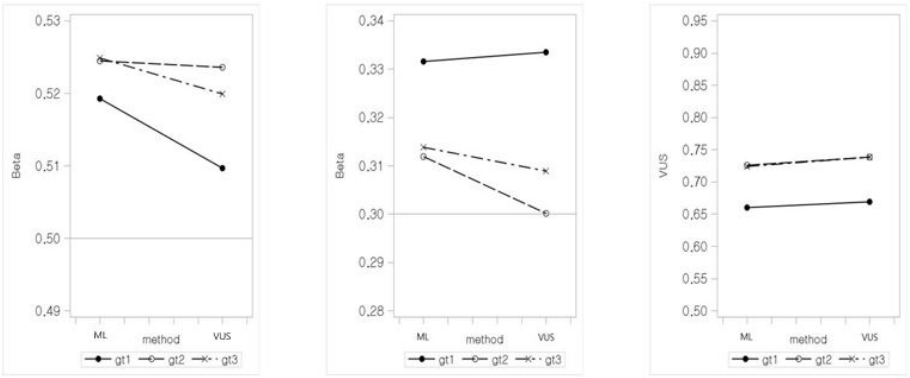

Figure 3.2 Coefficients and VUS for ML method and VUS approach ($\sigma = 0.2$)

Table 3.2는 분류점 c_1, c_2의 분포에서 표준편차의 크기를 두 배로 증가시킨 $N(0.3, 0.2^2)$과 $N(0.7, 0.2^2)$에서 생성한 뒤 추정량을 구한 결과이다. 연결함수가 $g_1(t)$인 경우 두 방법의 추정값의 평균은 크게 차이가 없었으나 연결함수가 $g_2(t)$, $g_3(t)$인 경우 ML 방법보다 VUS 접근방법의 추정값이 모수값인 0.5와 0.3에 가깝게 추정하는 결과를 보였다. 추정값의 편차 역시 연결함수가 $g_1(t)$인 경우 두 방법의 차이가 크지 않았지만 다른 두 연결 함수에 대하여 VUS 접근방법의 추정값의 표준편차가 ML 방법 추정값의 표준편차보다 작게 나와 안정적인 추정결과를 보인다. 분류점 c_1, c_2분포의 표준편차가 커지면

서 Table 3.1의 결과와 비교하여 추정값의 편차도 커지는 것을 확인할 수 있다.

VUS 접근방법으로 계산된 VUS는 ML 방법으로 계산된 VUS보다 큰 것을 알 수 있고, Table 3.1의 결과와 비교하여 전체적인 VUS가 작아진 것으로 보아 분류점의 분산이 커지면서 분류 성능은 앞의 결과보다 낮아진 것으로 탐색한다.

Figure 3.2의 β_2와 β_3의 추정값을 살펴보면, VUS 접근방법이 모수 0.5와 0.3에 ML 방법과 동등하거나 모수에 근접한 추정값을 나타내는 것을 파악할 수 있다. Figure 3.1과 비교하여 앞에서는 두 방법 모두 각 연결함수마다 편의는 있지만 추정값의 편차는 작았으나 Figure 3.2에서는 각 연결함수마다 추정값의 편차가 커지는 경향을 보였다. Figure 3.2의 VUS 추정값을 살펴보면, 앞의 Figure 3.1과 비교하여 전체적으로 VUS의 평균값이 작아졌지만 여전히 VUS 접근방법이 ML 방법보다 큰 값을 보였다.

Table 3.3 Coefficients and HUM for ML method and HUM approach ($\sigma = 0.1$)

Link	$\widehat{\beta}_2^{ML}$	$\widehat{\beta}_3^{ML}$	$\widehat{\beta}_2^{HUM}$	$\widehat{\beta}_3^{HUM}$	$\widehat{HUM}^{ML}$	$\widehat{HUM}^{HUM}$
$g_1(t)$	0.5256(0.1266)	0.3275(0.1090)	0.5105(0.0823)	0.3154(0.0765)	0.7206	0.7416
$g_2(t)$	0.5238(0.1355)	0.3121(0.1322)	0.5035(0.0644)	0.3003(0.0587)	0.7458	0.7887
$g_3(t)$	0.5125(0.1244)	0.3144(0.1259)	0.5099(0.0504)	0.3011(0.0621)	0.6978	0.7607

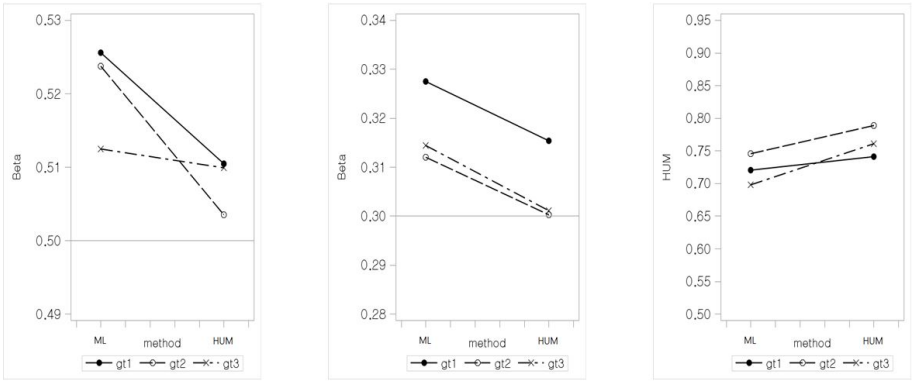

Figure 3.3 Coefficients and HUM for ML method and HUM approach ($\sigma = 0.1$)

Table 3.3과 Table 3.4는 HUM 접근방법을 이용한 모의실험의 결과로 앞의 VUS 접근방법의 모의실험과 같은 분포의 X들과 분류점 c_1, c_2 그리고 c_3를 각각 $N(0.25, 0.1^2)$, $N(0.5, 0.1^2)$, $N(0.75, 0.1^2)$에서 생성하고 다시 각각의 연결함수를 이용하여 $g(t) < c_1$이면 $Y = 1$, $c_1 \le g(t) < c_2$이면 $Y = 2$, $c_2 \le g(t) < c_3$이면 $Y = 3$, 그리고 $c_3 \le g(t)$이면 $Y = 4$로 설정한다. 각 연결 함수별 ML 방법과 HUM 접근방법으로 추정한 추정값의 평균을 나타냈다.

연결함수가 $g_1(t)$인 경우 $\widehat{\beta}_2^{ML}$, $\widehat{\beta}_3^{ML}$의 평균은 ML 방법의 추정값인 경우에 각각 0.5256, 0.3275였고, HUM 접근방법의 추정값인 경우에 각각 0.5105, 0.3154로 HUM 접근방법의 추정값이 편의가 작은 것으로 나타났다. 연결함수가 $g_2(t)$, $g_3(t)$인 경우도 HUM 접근방법의 모수 추정값이 로지스틱모형 접근방법보다 실제 모수값에 가까운 추정을 하는 것으로 나타났다. VUS 접근방법과 마찬가지로 연결함수가 로짓이 아닌 경우에 특히 HUM 접근방법이 좋은 추정 결과를 보여주는 것으로 나타났다. 효율성 측면에서 각각의 연결함수에 대하여 $\widehat{\beta}_2^{ML}$, $\widehat{\beta}_3^{ML}$의 표준편차가 각각 0.1266, 0.1355, 0.1244 그리고 0.1090, 0.1322, 0.1259으로 $\widehat{\beta}_2^{HUM}$과 $\widehat{\beta}_3^{HUM}$의 표준편차 0.0823, 0.0644, 0.0504 그리고 0.0765, 0.0587, 0.0621보다 큰 값을 보여 HUM 접근방법이 좋은 효율을 나타냈다.

$\widehat{HUM}^{HUM}$의 평균들이 0.7416, 0.7887, 07607으로 ML 방법의 $\widehat{HUM}^{ML}$보다 큰 값을 보이는 것을 확인하였다. 그러나 범주 $Y = 4$가 한 개가 늘어나면서 전반적인 HUM이 작아지는 것으로 보아 분류성능이 낮아지는 경향을 보였다.

Figure 3.3에서 β_2와 β_3의 추정값을 살펴보면 세 연결함수 모두의 경우에 대해 ML 방법은 HUM 접근방법보다 모수를 크게 추정하는 경향을 보여 HUM 접근방법이 원래의 모수 0.5와 0.3을 잘 추정하고 있는 것을 파악할 수 있다. HUM 추정값으로부터는 HUM 접근방법 추정값의 HUM이 최대화된 것을 확인할 수 있다.

Table 3.4 Coefficients and HUM for ML method and HUM approach ($\sigma = 0.2$)

Link	$\widehat{\beta}_2^{ML}$	$\widehat{\beta}_3^{ML}$	$\widehat{\beta}_2^{HUM}$	$\widehat{\beta}_3^{HUM}$	$\widehat{HUM}^{ML}$	$\widehat{HUM}^{HUM}$
$g_1(t)$	0.5231(0.1549)	0.3225(0.1566)	0.5225(0.1139)	0.3155(0.1192)	0.5462	0.5671
$g_2(t)$	0.5257(0.1768)	0.3149(0.1811)	0.5131(0.0952)	0.2956(0.1022)	0.5841	0.6284
$g_3(t)$	0.5112(0.1646)	0.3063(0.1419)	0.4985(0.0907)	0.2887(0.0642)	0.5376	0.5808

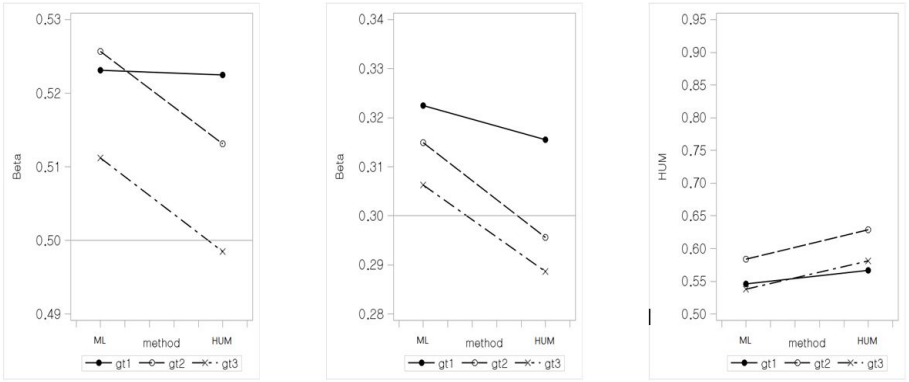

Figure 3.4 Coefficients and HUM for ML method and HUM approach ($\sigma = 0.2$)

Table 3.4는 분류점 c_1, c_2, c_3의 분포에서 표준편차가 두 배인 각각 $N(0.25, 0.2^2)$, $N(0.5, 0.2^2)$, $N(0.75, 0.2^2)$ 분포에서 생성한 뒤 추정한 결과로 VUS의 결과와 비슷하게 연결함수가 $g_1(t)$인 경우 두 방법의 추정값의 평균 차이가 크지 않았으나 연결함수가 $g_2(t)$, $g_3(t)$경우에 HUM 접근방법의 추정값인 $\widehat{\beta}_2^{HUM}$의 평균이 0.5131, 0.4985이고 $\widehat{\beta}_3^{HUM}$의 평균이 0.2956, 0.2887으로 ML 방법의 추정값인 $\widehat{\beta}_2^{ML}$, $\widehat{\beta}_3^{ML}$보다 비교적 모수를 잘 추정하는 것으로 나타났다.

각각의 연결함수에 대하여 $\widehat{\beta}_2^{HUM}$와 $\widehat{\beta}_3^{HUM}$의 표준편차가 0.1139, 0.0952, 0.0907 그리고 0.1192, 0.1022, 0.0642으로 $\widehat{\beta}_2^{ML}$, $\widehat{\beta}_3^{ML}$의 표준편차보다 비교적 작은 경향을 나타내 효율적인 결과를 보였다. HUM의 경우도 Table 3.3과 비교하여 분류점의 표준편차가 커지면서 두 방법 모두 추정값의 표준편차가 커지는 경향을 보였으나 HUM 접근방법이 여전히 ML 방법보다 안정적인 결과를 보였다.

분류성능측면에서 HUM 접근방법의 추정값으로 추정된 HUM의 평균이 0.5671, 0.6284, 0.5808으로 ML 방법보다 큰 HUM을 나타내었고 분류점의 편차가 커지면서 HUM들의 평균들이 작아지는 경향을 보여 분류성능도 낮아지는 것을 탐색할 수 있다.

Table 3.4의 결과에 대한 Figure 3.4에서도 세 연결함수 모두의 경우에 대해 HUM 접근방법이 ML 방법보다 모수를 잘 추정하였다. 앞의 VUS 모의실험 결과와 유사하게 Figure 3.3에서보다 Figure 3.4에서 연결함수마다 서로 추정값들의 편차가 큰 것을 파악할 수 있고 HUM이 두 방법 모두 Figure

3.3과 비교하여 작아졌지만 HUM 접근방법 HUM이 ML 방법의 HUM보다 여전히 큰 값을 나타내고 있다. 두 모의실험의 경우에서 ML 방법은 모수를 과추정 (overestimate)하는 경향을 보였으나 VUS와 HUM 접근방법은 이러한 경향을 보이지 않았다.

4. 결론

스코어가 선형결합으로 이루어진 형태의 ROC 곡선을 구하고 ROC 곡선의 평가측도인 AUC 통계량을 이용하여 선형결합모형의 모수를 추정하는 것은 매우 유용한 통계적 방법이다. Pepe (2005)는 AUC 접근방법이 로지스틱 가정이 맞지 않는 상황에서 좋은 추정 방법임을 제안하였고, Kraus (2014)는 일반적인 상황으로 확장하여 로짓 이외에 프로빗, complementary log-log 그리고 양수의 t값에서 빠르게 1로 수렴하도록 로짓을 변형한 함수를 고려하여 Pepe (2005)의 AUC 접근방법이 로지스틱모형의 최대가능도 추정량보다 좋은 결과를 보이는 것을 발견하였다.

본 연구에서는 세 가지 이상의 다항범주 판별결과를 고려하여 Pepe (2005)의 AUC 접근방법을 확장하였다. AUC 통계량의 정의를 바탕으로 VUS와 HUM 접근방법을 정의하고 두 방법을 Kraus (2014)가 고려한 다양한 연결함수에 적용하여 기존의 ML 방법과 비교한 결과를 얻었다.

판별결과가 세 가지 범주에 대한 VUS 접근방법은 연결함수가 로짓인 경우엔 ML 방법과 유사한 추정 결과를 보였고 그 외의 연결함수인 complementary log-log함수와 로짓을 변형한 함수에 대해서는 모수값에 가까운 추정 결과를 보였다. 동일한 연결함수에 대하여 두 분류점의 분포의 표준편차가 두 배 커졌을때, 두 방법 모두 모수값에 근접하는 추정값을 구하였으나 추정값의 표준편차가 커지고 VUS 값은 작아지는 경향을 보였다. 그럼에도 VUS 접근방법이 비교적 편의와 표준편차가 작았으므로 세 범주의 판별결과에 대하여 좋은 결과를 보였다. 네 범주의 판별결과에 대하여 동일하게 HUM 접근방법도 로짓 연결함수에 대해 ML 방법과 비슷한 결과를 보였고 그 외의 연결함수에 대하여 편의와 효율성 측면에서 좋은 결과를 보였다. HUM 접근방법도 세 분류점의 변동이 커지면서 모수 추정값의 표준편차가 커지고 HUM 값이 작아지는 경향을 보였지만 여전히 모수는 잘 추정하고 있었다. 즉 로지스틱 가정이 맞지 않는 다항범주 판별결과에 대해서도 VUS와 HUM 접근방법도 Pepe (2005)의 AUC 접근방법과 유사하게 ML 방법보다 동등하거나 더 나은 모수추정 결과를 보이는 것을 확인하였다.

References

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.

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. (2015). Two optimal threshold criteria for ROC analysis. *Journal of the Korean Data & Information Science Society*, **26**, 255-260.

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

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.

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.

Journal of the Korean Data & Information Science Society
2015, **26**(6), 1305–1315

http://dx.doi.org/10.7465/jkdi.2015.26.6.1305
한국데이터정보과학회지

Parameter estimation of linear function using VUS and HUM maximization

Chong Sun Hong[1] · Chi Hwan Won[2] · Dong Gil Jeong[3]

[123]Department of Statistics, Sungkyunkwan University

Received 3 August 2015, revised 8 September 2015, accepted 14 September 2015

Abstract

Consider the risk score which is a function of a linear score for the classification models. The AUC optimization method can be applied to estimate the coefficients of linear score. These estimates obtained by this AUC approach method are shown to be better than the maximum likelihood estimators using logistic models under the general situation which does not fit the logistic assumptions. In this work, the VUS and HUM approach methods are suggested by extending AUC approach method for more realistic discrimination and prediction worlds. Some simulation results are obtained with both various distributions of thresholds and three kinds of link functions such as logit, complementary log-log and modified logit functions. It is found that coefficient prediction results by using the VUS and HUM approach methods for multiple categorical classification are equivalent to or better than those by using logistic models with some link functions.

Keywords: Discrimination, link, manifold, risk, score, surface, threshold.

[1] Corresponding author: Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea. E-mail: cshong@skku.edu
[2] Graduate student, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea.
[3] Graduate student, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea.