응용통계연구 (2011)
24(2), 269–278
DOI: 10.5351/KJAS.2011.24.2.269

정규혼합분포를 이용한 ROC 분석

홍종선[1] · 이원용[2]

[1]성균관대학교 통계학과, [2]성균관대학교 통계학과

(2010년 12월 접수, 2011년 2월 채택)

요 약

스코어 변수의 민감도와 특이도와의 관계로 표현한 ROC 곡선을 더욱 정확한 진단을 위하여 분포함수와 공변량을 고려한 연구가 많이 진행되었다. 공변량을 고려하는 회귀분석 방법을 사용하였으며 이때 분포함수를 정규분포로 가정하거나 잔차의 분포함수를 추정하여 ROC 분석을 하였다. 본 연구는 분포함수가 주어지지 않으며 진단에 영향을 주는 공변량을 모르는 일반적인 상황에서 논의하였다. 확률변수인 스코어와 두 개의 부모집단으로 구성된 신용평가 자료에 적합한 분포함수를 추정하기 위하여 여러 개의 정규분포가 혼합된 정규혼합분포를 사용하여 ROC 분석을 한다. 고전적인 비모수적이고 경험적인 ROC 곡선에 적합한지를 파악하기 위하여 AUC 통계량을 사용하여 비교하며, 본 연구에서 제안한 정규혼합분포를 이용한 ROC 곡선이 다른 방법으로 구한 ROC 곡선보다 적합함을 보였다.

주요용어: 신용평가, 분류모형, 절단점, 준가능도함수.

1. 서론

ROC 곡선(Receiver Operating Characteristic Curve)은 성과(performance)를 기반으로 한 분류모형(classification model) 또는 분류자(classifiers)를 시각화할 수 있고 평가할 수 있는 유용한 방법이다. ROC 곡선은 분류자의 'hit rate'(이익) 또는 'sensitivity'(민감도)와 'false alarm rate'(비용) 또는 '1-specifity'(1-특이도) 사이에 교환(trade-off)을 나타내는 신호탐지이론에서 오랫동안 사용되어졌다 (Sobehart와 Keenan, 2001; Engelmann 등, 2003; Drummond와 Holte, 2006). 또한 의사결정과 의학진단의 체계에서 폭넓게 사용되어졌다 (Hanley와 McNeil, 1982; Swets, 1988; Zou, 2002). ROC 곡선의 특성에 관한 설명과 실증연구에서 ROC 분석을 응용하는데 관련된 정보는 Fawcett (2003)과 Provost와 Fawcett (1997, 2001), 홍종선과 최진수 (2009), 홍종선 등 (2010)에서 발견할 수 있다.

ROC 분석은 Swets와 Pickett (1982)의 고전적 교재의 출판이후 진단의 정확도를 묘사하는 의학의 응용분야에 보편화된 기술로 발전하였고, 최근에 ROC 분석은 통계적 연구에도 많은 성과가 있었다 (Gatsonis 등, 1995). Pepe (1998)는 진단 검정결과가 연속형이며, 연속형 진단결과의 정확도에 영향을 주는 공변량(covariate)을 평가하기 위하여 ROC 곡선에 대하여 회귀분석을 고려하였다. 일반적인 ROC 분석은 공변량을 고려하지 않은 자료를 기반으로 작성하여 비모수적 경험적(nonparametric empirical) ROC 분석이라고 한다. Pepe (2003)는 공변량을 고려하여 ROC 회귀분석(regression analysis)법을 제안하였는데, 스코어의 분포함수를 아는 경우에 선형모형을 고려한 모수적(parametric) 방법과 분포함수를 모르는 경우의 준모수적(semiparametric) 방법으로 구분하였다. Pepe가 제안한 두 가지의 방법으로

[1]교신저자: (110-745) 서울시 종로구 명륜동 3-53, 성균관대학교 경제학부 통계학전공, 교수.
E-mail: cshong@skku.ac.kr

구한 ROC 곡선이 고전적인 비모수적 경험적 ROC 곡선과의 적합을 연구하였다. 본 연구에서는 진단에 영향을 주는 공변량을 모르는 상황에서 스코어의 분포함수를 모르는 경우에 함수를 쉽게 그리고 가장 잘 적합시키는 방법 중의 하나로 정규혼합분포를 사용하여 분포함수를 추정하여 ROC 분석한다.

본 연구의 2절에서는 기존의 ROC 분석법들 중에서 세가지 방법을 간략히 정리한다. 고전적인 비모수적 경험적 ROC 곡선을 간략히 정리하며, Pepe (1998, 2003)가 제안한 모수적인 추정방법과 준모수적인 추정방법을 사용한 ROC 회귀분석법을 소개한다. 3절에서는 본 연구에서 제안한 정규혼합분포로 구성된 모형을 설명하고, 2절에서 소개한 기존의 방법들과 3절에서 제안한 방법으로 구한 ROC 곡선을 4절에서 비교분석한다. 마지막 5절에서는 본 연구에서 논의한 방법을 토론하면서 결론을 유도한다.

2. ROC 분석 방법

2.1. 비모수적 경험적 방법

본 연구에서는 진단 결과를 의학적 관점이 아닌 신용평가(credit evaluation)적 관점으로 논의하기 위하여 d(disease; 질병)와 n(non-disease; 정상)을 d(default; 부도)와 n(non-default; 정상)로 설정한다. 그리고 차주(borrower)는 스코어(score) 확률변수 S와 모수공간 D에 의해서 특성을 나타낸다고 가정하자. 확률변수 S는 대출기관에서 차주의 신용가치를 예상하기 위해 차주에게 부여한 연속형 값을 갖는 스코어이다. 스코어 변수 S를 통하여 대출기관은 궁극적으로 차주의 신용가치에 관한 정보에 의거하여 차주의 미래상태 D를 예상하는 것이다. 차주의 모집단은 두 개의 부모집단으로 구성된다고 가정한다. 즉 $D = \{d, n\}$이라고 하자. 부모집단은 미래시점에 대출상환능력이 없는 부도상태와 대출상환능력이 있는 정상상태로 구분된다. 차주의 모수 D가 d일 때($D = d$) 부도차주의 모집단에 속하고, 차주의 D가 n일 때($D = n$) 정상차주의 모집단에 속한다. 그리고 주어진 모수공간 D에서 스코어 변수의 조건부 누적분포함수를 각각 $F_d(\cdot) = P(S \le s \mid D = d)$와 $F_n(\cdot) = P(S \le s \mid D = n)$으로 나타낼 때, 스코어 S의 분포는 다음과 같이 표현할 수 있다.

$$F(s) = \gamma F_d(s) + (1 - \gamma) F_n(s), \tag{2.1}$$

여기서 γ는 전체부도율(total probability of default)이다. 즉 $\gamma = P[D = d]$이다.

ROC 곡선은 각 절단점(cut-off value, threshold)의 스코어에서 얻는 비율들로 구성되어 있으며, 실제 부도를 부도로 정확히 예측하는 비율 TPR(true positive rate)과 실제정상을 부도로 잘못 예측하는 비율 FPR(false positive rate)을 각각 Y축과 X축 좌표에 대응시킨 그래프로 다음과 같이 표현된다 (상세한 정보는 Tasche (2006) 참조).

$$\begin{aligned} &(F_n(s), F_d(s)), \quad s \in (-\infty, \infty) \\ &(u, \mathrm{ROC}(u)), \quad u \in (0, 1), \end{aligned}$$

여기서 $\mathrm{ROC}(u) = F_d(F_n^{-1}(u))$, $u \in (0, 1)$이다.

2.2. 회귀모형을 이용한 모수적 방법

스코어의 분포함수를 알고 있다는 가정 하에서 ROC 분석을 논의하기 위하여 식 (2.1)의 분포를 다음과 같이 정규분포로 가정하여 살펴보자. 특정한 절단점 c에 대하여

$$F_d(c) = \Phi\left(\frac{c - \mu_d}{\sigma_d}\right), \qquad F_n(c) = \Phi\left(\frac{c - \mu_n}{\sigma_n}\right).$$

$u = \Phi((c-\mu_n)/\sigma_n)$로 설정하면, $c = \mu_n + \sigma_n \Phi^{-1}(u)$이므로 ROC$(u)$는 이봉정규(binormal) 형태로 표현된다.

$$\Phi\left(a + b\Phi^{-1}(u)\right),$$

여기서 $a = (\mu_n - \mu_d)/\sigma_d$, $b = \sigma_n/\sigma_d$이다.

다음으로는 ROC 곡선의 정확도를 높이기 위해 공변량을 고려한 회귀분석을 이용하기 위하여 공변량 Z가 있을 때 위치-모수모형(location-scale model)은 다음과 같이 가정한다 (Pepe, 1998, 2003).

$$S = \mu(D, Z) + \sigma(D)\varepsilon,$$

여기서 Z는 공변량 (X_1, X_2)이고, ε는 평균이 0이고 분산이 1인 분포함수를 $F_0(\cdot)$를 따른다고 설정하자. 지시함수 $I_D[d]$=1에 대하여 평균함수를

$$\mu(D, Z) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 I(D) + \alpha_4 I(D) X_1 + \alpha_5 I(D) X_2$$

그리고 분산함수는 공변량에 의존하지 않는

$$\sigma(D) = \sigma_d^{I[D]} \sigma_n^{1-I[D]}$$

로 가정하면, ROC 함수는 다음과 같다.

$$\begin{aligned} \mathrm{ROC}(u) &= F_0\left[\left\{\frac{\mu(n,Z) - \mu(d,Z)}{\sigma_d}\right\} + \left(\frac{\sigma_n}{\sigma_d}\right) F_0^{-1}(u)\right] \\ &= F_0\left(a_0 + a_1\ X_1 + a_2 X_2 + b F_0^{-1}(u)\right), \end{aligned} \tag{2.2}$$

여기서 $a_0 = -\alpha_3/\sigma_d$, $a_1 = -\alpha_4/\sigma_d$, $a_2 = -\alpha_5/\sigma_d$, $b = \sigma_n/\sigma_d$이다. 분포함수 F_0를 표준정규분포를 대체한다면 다음과 같이 이봉정규 형태로 표현된다.

$$\mathrm{ROC}\,(u) = \Phi\left(a_0 + a_1 X_1 + a_2 X_2 + b\Phi^{-1}(u)\right).$$

2.3. 회귀모형을 이용한 준모수적 방법

2.2절에서와 달리 분포함수 F_0를 모른다고 가정하는 경우에는 $\underline{\alpha}$의 최대가능도 추정량을 구할 수 없으므로 표준화 잔차의 분포를 이용하여 준가능도(Quasi-likelihood) 방법을 사용한다. 평균함수와 분산함수를 2.2절에서와 같이 동일하게 가정하고 추정량 $\underline{\alpha} = (\alpha_0, \alpha_1, \alpha_2)'$와 $\sigma(D)$는 다음을 만족한다 (McCullagh와 Nelder, 1983 참조).

$$\sum \left\{\left(\frac{\partial}{\partial \underline{\alpha}}\right)\mu(D,Z)\right\}\left(\frac{S - \mu(D,Z)}{\sigma(D)}\right) = 0,$$

여기서 $\sigma(D)$의 추정량은 $\hat{\sigma_d^2} = \sum^{N_d}(S_d - \hat{\mu}(d,Z))^2/N_d$, $\hat{\sigma_n^2} = \sum^{N_n}(S_n - \hat{\mu}(n,Z))^2/N_n$이다. 분포함수 F_0를 추정하기 위해 표준화 잔차 $\{(S_i - \hat{\mu}(D,Z))/\hat{\sigma}_i(D),\ i = 1, \ldots, N = N_n + N_d\}$를 이용하여 다음과 같이 추정한다.

$$\hat{F_0}(s) = N^{-1}\sum_{i=1}^{N} I\left[\frac{S_i - \hat{\mu}_i(D,Z)}{\hat{\sigma}_i(D)} \le s\right].$$

추정된 분포함수로 표현된 ROC 함수는 다음과 같다.

$$\widehat{\mathrm{ROC}}(u) = \hat{F_0}\left(\hat{a_0} + \hat{a_1} X_1 + \hat{a_2} X2 + \hat{b}\hat{F_0}^{-1}(u)\right). \tag{2.3}$$

2.2절에서는 F_0를 표준정규분포로 가정했으나 2.3절에서는 F_0를 추정하는 점만 다를 뿐 준가능도함수에서 구한 모수는 최대가능도 추정량과 동일하다 (Pepe, 2003, pp. 148).

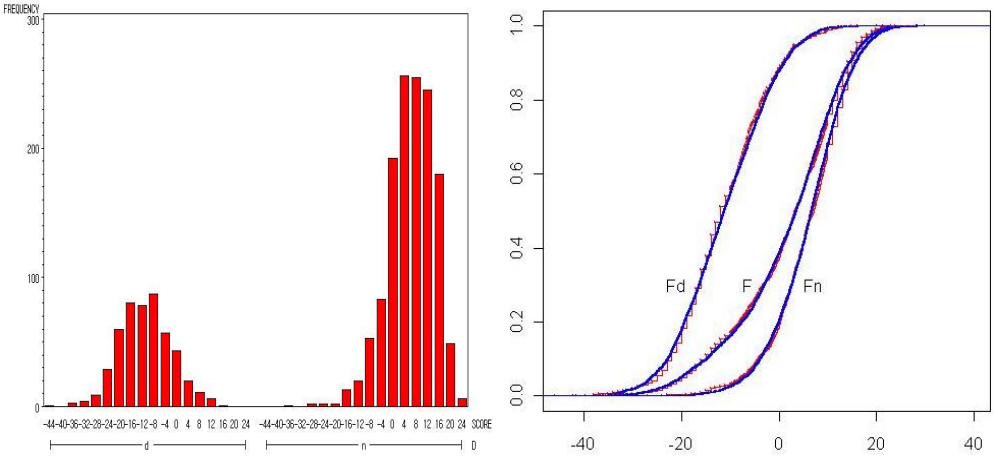

그림 **3.1.** DP21 자료의 분포와 누적분포함수

3. 정규혼합 방법

현실적인 상황으로 식 (2.1)의 분포함수를 모르고 평균함수에 영향을 주는 공변량도 고려하지 않는다고 가정하자. 이런 상황에서 본 연구는 자료에 가장 적합한 분포함수를 추정하는 방법 중에서 편리한 방법인 정규혼합(normal mixture) 분포를 이용한다. 분포함수 $F_d(s)$와 $F_n(s)$를 각각 p개와 q개의 정규분포함수의 선형결합(linear combination)으로 구성되었다고 하자.

$$F_d(s) = \sum_{i=1}^{p} \alpha_i \Phi\left(s; \mu_{d_i}, \sigma_{d_i}^2\right), \qquad F_n(s) = \sum_{j=1}^{q} \beta_j \Phi\left(s; \mu_{n_j}, \sigma_{n_j}^2\right),$$

여기서 $\sum_{i=1}^{p} \alpha_i = 1$, $\sum_{j=1}^{q} \beta_j = 1$ 그리고 $\Phi(s; \mu, \sigma^2)$는 평균과 분산이 각각 μ, σ^2인 정규분포함수를 나타낸다. 정규혼합분포를 이용하여 분포함수를 추정하고, 이 함수를 사용하여 ROC 분석을 한다.

다음 절에서는 비모수적 경험적 방법과 회귀분석을 이용한 모수적 그리고 준모수적인 방법인 기존의 ROC 분석방법들과 본 연구에서 제안한 정규혼합분포를 이용하는 방법을 비교하기 위하여 두 종류의 예제에 대하여 실증분석한다. 우선 4절에서 논의할 두 실증예제 자료에 적합한 정규혼합분포를 구하여보자. 첫 번째로 Pepe (2003)의 연구에 사용된 자료는 1848명의 사람들 중 489명의 사람은 청각장애가 있고 1359명의 사람은 청각장애가 없는 정상인의 자료(이하 DP21)이다 (Stover 등, 1996). 이 자료에 적합한 정규혼합분포를 추정한 결과는 다음과 같다. $\gamma = 489/1848$이며,

$$F_d(s) = 0.6\Phi(s|\,-16.1, 7.2) + 0.4\Phi(s|\,-3.9, 6.6),$$
$$F_n(s) = 0.4\Phi(s|1.54, 7.55) + 0.6\Phi(s|\,9.32, 5.78).$$

DP21 자료의 경험적 누적분포함수에 대하여 정규혼합분포의 적합도를 파악하는 콜모고로프-스미르노프(Kolmogorov-Smirnov) 검정 결과 통계량값은 0.0282이며 이에 대응하는 p-값은 0.2337로 매우 큰 값으로 나타났다. 따라서 정규혼합을 이용한 분포가 실제 분포에 매우 적합하다고 판단된다.

그림 3.1은 DP21 자료의 분포를 표현한 것으로 왼쪽 그림은 부도(분포의 왼쪽부분)와 정상(분포의 오른쪽부분)으로 구분하여 막대그래프(histogram)로 나타내었다. 오른쪽 그림은 부도와 정상 그리고 전체의 자료 각각에 대응하는 경험적 누적분포함수(empirical cumulative distribution function)와 추정

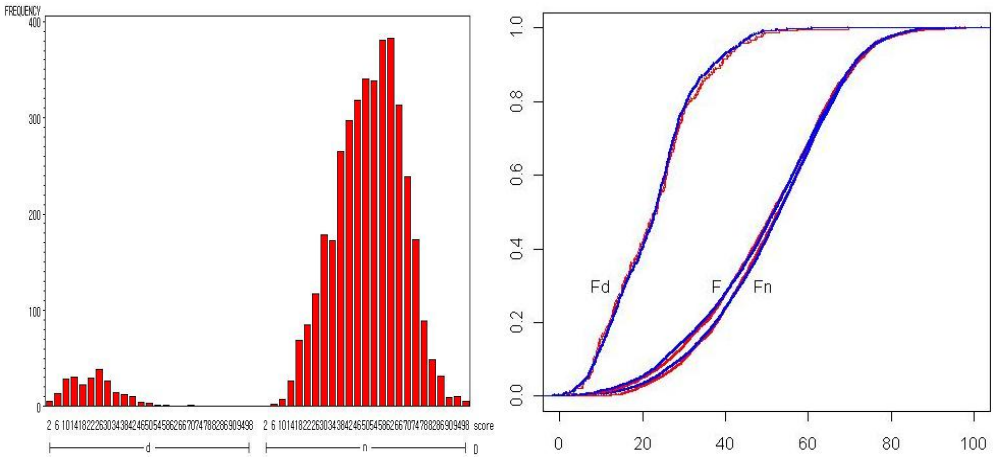

그림 3.2. 외감기업 자료의 분포와 누적분포함수

된 정규혼합 분포함수를 나타내었다. 경험적 누적분포함수는 계단식 선으로 표현하고, 추정된 정규혼합 분포함수는 실선으로 표현하였다. 세 종류의 실선이 대응하는 계단식 선 위에 거의 겹치고 있으며 이것은 추정된 정규혼합 분포함수가 경험적 누적분포함수에 적합하다고 판단할 수 있다.

두 번째 자료는 1994년부터 2005년까지 한국 기업 중에서 외부 감사를 받는 기업 중 총자산 규모가 4500억원 이상인 기업에 대한 자료(이하 외감자료)이며, 총 표본수는 4,134 ($N_d = 238$, $N_n = 3{,}896$)이다. 이 자료에 적합한 정규혼합분포를 추정한 결과는 다음과 같다. $\gamma = 238/4134$이며,

$$F_d(s) = 0.4\Phi(s|12.5, 4.8) + 0.3\Phi(s|26, 3.25) + 0.3\Phi(s|33.65, 9.05),$$
$$F_n(s) = 0.4\Phi(s|38.5, 13.3) + 0.6\Phi(s|60.5, 11.3).$$

이 정규혼합분포의 적합도 검정통계량인 콜모고로프-스미르노프 통계량값은 0.0141이며 이에 대응하는 p-값은 0.7609로 매우 큰 값으로 나타났다. 따라서 정규혼합을 이용한 분포가 실제 분포에 매우 적합하다고 판단된다.

그림 3.2의 왼쪽 그림은 외감기업 자료의 부도(분포의 왼쪽부분)와 정상(분포의 오른쪽부분)으로 구분하여 막대그래프로 표현하였으며, 오른쪽 그림은 부도와 정상 그리고 전체의 자료 각각에 대응하는 경험적 누적분포함수(계단식 선)와 추정된 정규혼합 분포함수(실선)를 나타내었다. 그림 3.2를 통해서 세 종류의 실선이 대응하는 계단식 선에 모두 일치하고 있으며 이것은 추정된 정규혼합 분포함수가 경험적 누적분포함수에 적합하고 따라서 추정된 정규혼합 분포함수가 자료를 대표하는 분포함수라고 판단할 수 있다.

4. 실증예제

4.1. 예제1

3절에서 소개한 청각장애의 유무에 관한 자료(DP21)로, 공변량으로 소리 강도(intensive)와 소리 주파수(frequency)를 채택하였다. Pepe (2003)는 비모수적 경험적 ROC 곡선과 회귀분석을 이용한 모수적 그리고 준모수적인 방법인 기존의 ROC 곡선을 비교분석하였다 (모형의 결과는 Pepe (2003) 참조). 회귀분석을 이용한 모수적 방법으로 구한 모수의 추정량은 표 4.1과 같으며, 준모수적 방법에서도

표 **4.1.** DP21 자료의 모수추정

공변량	모수	추정값
상수	α_0	1.08630
음성주파수	α_1	−0.14592
음성강도	α_2	−0.85894
청각장애	α_3	29.19587
청각장애×음성주파수	α_4	0.46700
청각장애×음성강도	α_5	−3.14719
모수	σ_d	8.97609
	σ_n	7.76571

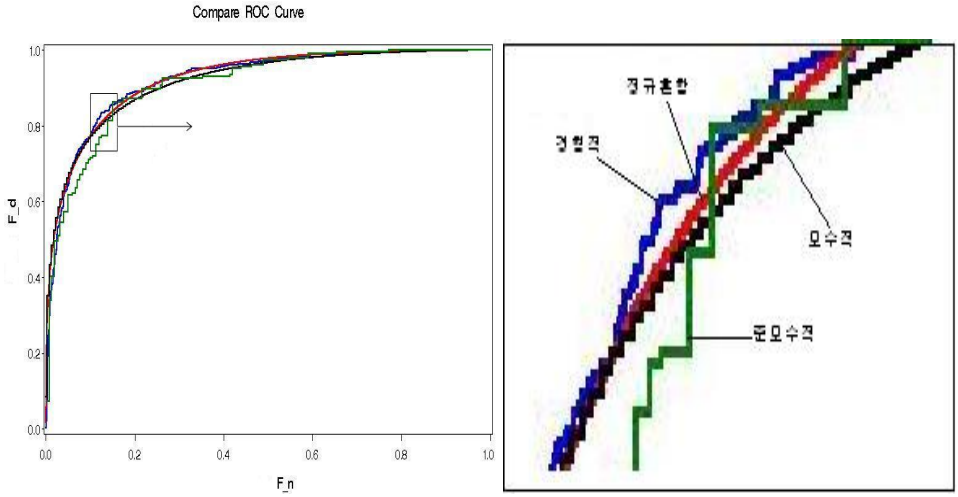

그림 **4.1.** DP21 자료의 ROC 곡선

표 4.1의 추정량을 사용하며, 다만 표준화된 잔차를 이용하여 경험적 누적분포함수를 분포함수 F_0의 추정분포로 가정하여 ROC 함수를 구하였다.

세 종류의 기존 방법과 정규혼합을 이용하여 구한 ROC 곡선을 겹쳐서 구하여 그림 4.1에 표현하였다. 그림 4.1은 경험적 방법을 이용한 ROC 곡선과 모수적 방법을 이용한 ROC 곡선과 준모수적 방법을 이용한 ROC 곡선과 정규혼합 방법을 이용한 ROC 곡선을 표현하였고 ROC 곡선의 일부를 확대하여 그림 4.1의 오른쪽에 나타내었다. 경험적 방법을 이용한 ROC 곡선에 가장 근접하는 것은 정규혼합을 이용한 ROC 곡선이며 준모수적 방법을 이용한 ROC 곡선이 가장 멀리 떨어져 있는 것을 파악할 수 있다. 따라서 정규혼합 방법을 이용한 ROC 곡선이 비모수적 경험적 ROC 곡선에 가장 적합함을 알 수 있었다. 비모수적 경험적 ROC 곡선과 비교하여 얼마나 근사한지를 판단하기 위하여

각각의 ROC 곡선의 AUC(area under curve)를 구하여 비교하였다. 모형의 변별력이 얼마나 정확한가를 평가하는 AUC 통계량은 ROC 곡선 아래의 면적을 측정하는 척도로 다음과 같이 정의하고, 그 결과를 표 4.2에 나타내었다.

$$\text{AUC} = 1 - \sum_i f_d(s_i) \times \left[\frac{F_n(s_{i-1}) + F_n(s_i)}{2} \right],$$

표 **4.2.** DP21 자료의 AUC 통계량

모형	경험적 방법	모수적 방법	준모수적 방법	정규혼합 방법
AUC 통계량	0.92289	0.91825	0.91142	0.92521
(경험적 방법과 차이)		(0.00464)	(0.01147)	(0.00232)

표 **4.3.** 외감기업 자료의 모수추정

공변량	모수	추정값
상수	α_0	-31.07612
자기자본비율	α_1	9.59785
부도여부	α_2	-0.48852
부도여부×자기자본비율	α_3	0.18005
모수	σ_d	8.25737
	σ_n	10.26000

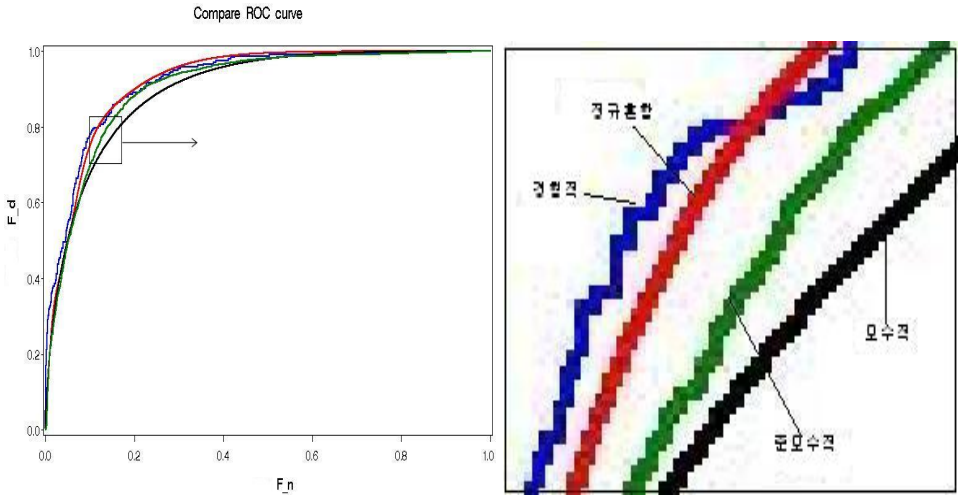

그림 **4.2.** 외감기업 자료의 ROC 곡선

여기서 $F_n(s_0) = 0$이다.

네 가지 방법을 이용하여 기준이 되는 비모수적 경험적 ROC 곡선과 비교한 표 4.2의 결과를 살펴보면, 정규혼합을 이용한 ROC 곡선이 가장 적합하다.

4.2. 예제2

3절에서 논의한 외감기업 자료로, 공변량으로 각 기업의 자기자본비율을 고려하였다. 모수적 방법으로 구한 모수의 추정량은 표 4.3과 같으며, 준모수적 방법에서 사용한 F_0의 추정분포는 표준화된 잔차를 이용하여 구한 경험적 누적분포함수를 사용하였다.

정규혼합을 이용하여 구한 ROC 곡선과 기존의 방법으로 구한 곡선과 함께 그림 4.2에 표현하였다. 그림 4.2에서도 경험적 방법을 이용한 ROC 곡선 모수적 방법을 이용한 ROC 곡선과 준모수적 방법을 이용한 ROC 곡선과 정규혼합 방법을 이용한 ROC 곡선을 표현하였고 ROC 곡선의 일부를 확대하여 그

표 **4.4.** 외감기업 자료의 AUC 통계량

모형	경험적 방법	모수적 방법	준모수적 방법	정규혼합 방법
AUC 통계량	0.92273	0.90013	0.90634	0.92123
(경험적 방법과 차이)		(0.00226)	(0.01639)	(0.0015)

림 4.2의 오른쪽에 나타내었다. 경험적 방법을 이용한 ROC 곡선에 가장 근접하는 것은 정규혼합을 이용한 ROC 곡선이며 모수적 방법을 이용한 ROC 곡선이 가장 멀리 떨어져 있는 것을 파악할 수 있다. 그러므로 비모수적 경험적 ROC 곡선과 정규혼합 방법을 이용한 ROC 곡선이 가장 적합함을 알 수 있었다.

비모수적 경험적 ROC 곡선과 비교하여 근사한지를 판단하기 위하여 ROC 곡선의 AUC를 구하여 비교한 표 4.4를 살펴보면, 4.1절에서 논의한 것과 동일하게 정규혼합을 이용한 ROC 곡선이 가장 적합함을 파악할 수 있다.

5. 결론

스코어의 분포함수와 이에 대응하는 적절한 공변량을 고려하지 않고 스코어 자료만으로 분석한 ROC 곡선을 더욱 정확한 진단을 위하여 분포함수를 가정하거나 추정하고 또는 공변량까지 고려한 다양한 분석방법들이 개발되었다. Pepe (1998, 2003)는 연속형 진단결과를 가진 자료의 정확도에 영향을 주는 공변량을 고려하여 분포함수를 아는 경우와 모르는 경우의 두 종류의 회귀분석법을 제안하였다. 특히 분포함수가 주어진 경우는 정규분포로 가정하면서 분석하였고, 분포함수가 주어지지 않은 경우에는 함수를 추정하여 분석하였다.

본 연구에서는 일반적으로 분포함수가 주어지지 않은 경우를 고려하였으며, 자료에 적합한 함수를 추정하기 위하여 정규혼합 분포함수를 사용하여 ROC 분석을 하였다. Pepe가 제안한 두 종류의 ROC 곡선 그리고 본 연구에서 제안한 정규혼합 분포함수를 이용한 ROC 곡선이 고전적인 비모수적 경험적인 ROC 곡선을 적절하게 표현하는지를 판단하기 위하여 자료 수가 많은 두 가지의 예제로 실증분석하였다. 또한, 고전적인 ROC 곡선에 적합한지를 AUC 통계량을 사용하여 비교하였다. 그 결과 본 연구에서 제안한 정규혼합을 이용한 ROC 곡선이 고전적인 ROC 곡선에 가장 적합하다고 판단되었다. 그러므로 연구대상이 자료를 ROC 곡선으로 구현할 때 본 연구에서 제안한 혼합분포에 적합한 정규혼합분포를 추정하여 이용한다면, 자료를 잘 설명하는 ROC 분석에 활용할 수 있다.

참고문헌

홍종선, 주재선, 최진수 (2010). 혼합분포에서의 최적분류점, <응용통계연구>, **23**, 13–28.

홍종선, 최진수 (2009). ROC와 CAP 곡선에서의 최적분류점, <응용통계연구>, **22**, 911–921.

Drummond, C. and Holte, R. C. (2006). Cost curves: An improved method for visualizing classifier performance, *Machine Learning*, **65**, 95–130.

Engelmann, B., Hayden, E. and Tasche, D. (2003). Measuring the discriminative power of rating systems, *Discussion Paper, Series 2: Banking and Financial Supervision*.

Fawcett, T. (2003). ROC Graphs: Notes and practical considerations for data mining researchers, *Technical Report HPL-2003-4*, HP Laboratories, 1–28.

Gatsonis, C. A., Begg, C. B. and Wieand, S. A. (1995). Introduction to advances in statistical methods for diagnostic radiology: A symposium, *Academic Radiology*, **2**, S1–3.

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

McCullagh, P. and Nelder, J. A. (1983). Quasi-likelihood functions, *Annals of Statistics*, **11**, 59–67.

Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results, *Biometrics*, **54**, 124–135.

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

Provost, F. and Fawcett, T. (1997). Analysis and visualization of classifier performance comparison under imprecise class and cost distributions, In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, 43–48.

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

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

Stover, L., Gorga, M. P. and Neely, T. (1996). Towards optimizing the clinical utility of distortion product otoacoustic emission measurements, *Journal of the Acoustical Society of America*, **100**, 956–967.

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

Swets, J. A. and Pickett, R. M. (1982). *Evaluation Diagnostic Systems, Methods from Signal Detection Theory*, Academic Press, New York.

Tasche, D. (2006). Validation of internal rating systems and PD estimates, On-line bibliography available from: http://arXiv:physics/0606071.

Zou, K. H. (2002). Receiver operating characteristic literature research, On-line bibliography available from: http://www.spl.havard.edu/pages/ppl/zou/roc.html.

ROC Curve Fitting with Normal Mixtures

Chong Sun Hong[1] · Won Yong Lee[2]

[1]Department of Statistics, Sungkyunkwan University
[2]Research Institute of Applied Statistics, Sungkyunkwan University

(Received December 2010; accepted February 2011)

Abstract

There are many researches that have considered the distribution functions and appropriate covariates corresponding to the scores in order to improve the accuracy of a diagnostic test, including the ROC curve that is represented with the relations of the sensitivity and the speci?city. The ROC analysis was used by the regression model including some covariates under the assumptions that its distribution function is known or estimable. In this work, we consider a general situation that both the distribution function and the e?ects of covariates are unknown. For the ROC analysis, the mixtures of normal distributions are used to estimate the distribution function ?tted to the credit evaluation data that is consisted of the score random variable and two sub-populations of parameters. The AUC measure is explored to compare with the nonparametric and empirical ROC curve. We conclude that the method using normal mixtures is ?tted to the classical one better than other methods.

Keywords: Classification model, credit evaluation, quasi-likelihood, threshold.

[1]Corresponding author: Professor, Department of Statistics, Sungkyunkwan University, 3-53, Myungryun-Dong, Jongro-Gu, Seoul 110-745, Korea. E-mail: cshong@skku.ac.kr