한국통계학회논문집
2009, 16권, 2호, 239-247

신용평가모형에서 타당성검증 통계량들의 판단기준

박용석[a], 홍종선[1,b], 임한승[c]

[a]성균관대학교 응용통계연구소, [b]성균관대학교선 통계학과, [c]한국기업평가

요 약

신용평가모형의 판별력에 대한 검정방법으로 콜모고로프-스미르노프, 평균차이, AUROC, AR 등과 같은 통계량이 널리 사용되고 있다. 이러한 통계량들의 판단기준은 정규분포 가정 하에서 평균차이를 기준으로 설정되었다. 본 연구에서는 모의실험을 통해서 표본크기, 불량률 그리고 제II종 오류율을 고려하는 대안적인 판단기준을 제안하고 현재 적용되고 있는 판단기준과 비교해본다. 또한 판별력 정도에 따른 각 통계량들의 의미를 10단계로 정의하고 모의실험 결과와 현재 적용되고 있는 판단기준을 비교해 본다.

주요용어: 오분류율, 판별력, 평균차이, AR, AUROC.

1. 서론

ROC(receiver operating characteristic) 곡선과 CAP(cumulative accuracy profile) 곡선은 신용평가 모형의 성능(performance)을 탐색하는 유용한 시각적인 방법이다. ROC 곡선은 모형의 정분류율과 오분류율의 변화를 시각적으로 나타내기 위해서 이용되어왔으며 신용평가 분야와 같이 사전에 불량거래자를 정확하게 판단해야하는 모형의 판별력에 대한 시각적인 방법으로 확장되었다 (Egan, 1975; Swet, 1988; Swets 등, 2000). 이러한 시각적인 방법에 의존하는 것은 평가기준으로써 주관적이기 때문에 ROC 곡선을 수치화하여 나타낸 것을 ROC 측도(ROC measure)라 하고 이 측도는 ROC 곡선 아래의 면적을 계산하기 때문에 AUROC(area under ROC)라고 한다.

CAP 곡선은 ROC 곡선과 마찬가지로 신용평가 모형의 성능을 탐색하는 시각적인 방법이다 (Sobehart 등, 2000; Tasche, 2006). CAP 곡선을 수치화하여 나타낸 것이 AR(accuracy ratio)이고 이는 완전한 모형(perfect model)의 면적과 실제 구축된 평가 모형(rating model)의 면적 사이의 비율을 나타낸다. AUROC와 AR은 신용평가 분야에서 가장 널리 이용되는 모형의 판별력 검정 통계량으로 자세한 내용은 2절에서 살펴보기로 한다.

AUROC와 AR은 1에 가까우면 좋다는 상대적인 판단기준이 경험적으로 존재하며 두 개 또는 그 이상의 신용평가 모형이 존재하는 경우 어느 모형의 성능이 가장 좋은가에 대한 상대적인 판단기준으로 적용하고 있다. 하지만 K-S 통계량의 통계적 임계값과 같은 절대적인 판단기준은 존재하지 않는다. Joseph (2005)는 AUROC와 AR에 대한 판별력 판단기준으로 평균차이(mean difference)에 대응하는 기준을 제안하였다. Joseph (2005)가 제안한 판별력 판단기준은 Wilkie (2004)가 제안한 방법을 확장한 것으로 불량과 정상이 동일한 표준편차를 갖는 정규분포 가정 하에서 평균차이에 대응하는 AUROC와 AR의 판단기준 값을 제안하였다. Joseph (2005)가 제안한 판단기준은 표 1과 같다.

확률변수 X를 스코어(score)라고 하고 스코어값 x의 범위는 $(-\infty, \infty)$이다. 스코어에 대응하는 '정상(goods)'과 '불량(bads)'의 누적분포함수는 각각 $F_G(x)$, $F_B(x)$이고 이에 대응하는 확률밀도함수는 각각 $f_G(x)$, $f_B(x)$라 하자. 그러면 '정상'과 '불량'의 평균은 각각 μ_G, μ_B이고 표준편차는 동일하다고 가

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

표 1: 정규분포 가정에 기초한 통계량들의 판단기준

의미	MD	K-S	AR	AUROC
Random	0.00	0.00	0.00	0.50
Doubtfull	0.25	0.10	0.14	0.57
Poor	0.50	0.20	0.28	0.64
Marginal	0.75	0.29	0.40	0.70
Satisfactory	1.00	0.38	0.52	0.76
Good	1.25	0.47	0.62	0.81
Very Good	1.50	0.55	0.71	0.86
Strong	1.75	0.62	0.78	0.89
Very Strong	2.00	0.68	0.84	0.92
Excellent	2.25	0.74	0.90	0.95
Excellent	2.50	0.79	0.94	0.97
Excellent	2.75	0.83	0.97	0.99
Superior	3.00	0.87	0.99	1.00

정하여 σ라 하면, 표 1의 평균차이(MD)는 $(\mu_G - \mu_B)/\sigma$로 나타낼 수 있다 (Wilkie, 2004). 동일한 가정하에서 AUROC는 다음과 같이 구할 수 있다 (Faraggi와 Reiser, 2002).

$$\text{AUROC} = \Phi\left(\frac{\mu_G - \mu_B}{\sqrt{2}\sigma}\right). \tag{1.1}$$

AR은 AUROC와 선형관계를 갖으며 AR에 대한 기준값을 간단히 구할 수 있다 (Engelmann 등, 2003).

$$\text{AR} = 2\text{AUROC} - 1. \tag{1.2}$$

표 1의 판단기준은 두 분포가 정규분포이고 표준편차가 동일하다는 가정 하에서 평균차이를 기준으로 설정된 판단기준이다. 하지만 일반적인 신용평가 모형을 위해 수집된 자료는 정상과 불량의 표본크기의 차이가 크기 때문에 두 집단의 표준편차가 같다는 가정을 만족하기가 어려워진다.

박용석과 홍종선 (2008)은 표 1의 K-S 통계량에 대한 판별력 판단기준에 대해서 표본크기, 불량률 그리고 제II종 오류율의 차이를 고려한 대안적인 판단기준을 제안하고 제안된 판단기준을 기반으로 한 가설검정 방법을 제시하였다. 대안적인 판단기준은 표본크기, 불량률 그리고 제II종 오류율에 따라서 기준을 제시한다. 본 연구에서는 박용석과 홍종선 (2008)의 연구방법을 적용하여 K-S 통계량에 추가적으로 평균차이(MD), AUROC, AR 통계량에 대한 대안적인 판단기준을 제시한다. 또한 판별력 정도에 대응하는 판별력 판단을 위한 의미를 제안한다. 마지막으로 기존에 적용되고 있는 표 1의 판단기준과 대안적인 판단기준을 비교함으로써 표본크기의 차이가 발생하는 경우 통계량의 특징을 비교한다.

본 논문의 구성은 다음과 같다. 2절에서는 판별력 검정통계량인 AUROC와 AR에 대해서 설명한다. 3절에서는 모의실험 절차에 대해서 설명하고 모의실험 결과를 통해 판별력 검정 통계량의 분포를 살펴보고 표본크기, 불량률 그리고 제II종 오류율에 대응하는 판단기준을 제시하며 사례를 통한 대안적 판단기준의 적용방법을 살펴본다. 추가적으로 통계량의 판별력 정도에 따라서 보다 해석하기 쉬운 의미를 통해 판별력 정도를 판단하는 방법을 논의하며 기존에 적용되고 있는 통계량들의 판단기준과 대안적인 판단기준을 비교한다. 마지막으로 4절에서는 모의실험 결과에 대해서 정리하고 토론한다.

2. AUROC와 AR

X와 Y를 각각 불량과 정상의 스코어를 나타내는 서로 독립인 확률변수라고 하자. ROC 곡선 아래의 면적인 AUROC는 다음과 같이 계산된다 (Tasche, 2006).

$$\text{AUROC} = P(X < Y). \tag{2.1}$$

불량과 정상에 대응하는 스코어값을 각각 $x_1, \ldots, x_n$, $y_1, \ldots, y_m$이라 할 때 식 (2.1)은 다음과 같은 Wilcoxon-Mann-Whitney 통계량의 값과 같이 추정한다 (Hanley와 McNeil, 1982).

$$\widehat{\text{AUROC}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[I(x_i < y_j) + \frac{1}{2} \times I(x_i = y_j)\right]. \tag{2.2}$$

여기서 n은 불량의 개수이고 m은 정상의 개수이다. AUROC의 범위는 $0.5 \leq \text{AUROC} \leq 1.0$이고 1에 가까울수록 판별력이 좋은 모형이다. AR 통계량은 다음과 같이 정의되며 식 (2.4)에 의해 쉽게 구할 수 있다.

$$\text{AR} = P(X < Y) - P(X > Y). \tag{2.3}$$

$$\widehat{\text{AR}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[I(x_i < y_j) - I(x_i > y_j)\right]. \tag{2.4}$$

AR의 범위는 $0 \leq \text{AR} \leq 1$이고 1에 가까울수록 판별력이 좋은 모형이다.

3. 모의실험

3.1. 모의실험을 위한 가정

판별력 검정 통계량에 대한 모의실험을 위해서는 서로 다른 판별력에 대응하는 스코어의 생성이 필요하다. 본 연구에서는 다양한 판별력을 갖는 모형의 생성을 위해서 표본불량률을 정의한다. 모형 설정을 위해 수집된 자료 중에서 불량과 정상이 차지하는 비율은 사전에 알고 있다고 가정하자. 정상이 0, 불량이 1의 값을 갖는 지시변수(indicator variable) Z로 정의하면 전체불량률(total probability of bads)은 $p = P(Z = 1) = 1 - P(Z = 0)$로 정의할 수 있고 전체자료를 N개라 하면 불량의 개수는 $n \approx Np$이고 정상의 개수는 $m = N - n$으로 표현된다. 그리고 $F(x_{(n)}) = r$을 만족하는 r을 표본불량률이라고 정의하면 표본불량률과 전체불량률은 일반적으로 $r > p$의 관계가 성립한다. 표본불량률 $\hat{r}$은 다음과 같다.

$$\hat{r} = \left[n + \sum_{j=1}^{n} I(Y_j \leq x_{(n)})\right] \Big/ N. \tag{3.1}$$

모형에 대한 판별력 판단을 위한 다른 방법으로 오분류율(misclassification rate)을 고려할 수 있다. 오분류율은 모형에서 정상과 불량의 분류기준 스코어(cut-off score)에 의해서 달라진다. 분류기준 스코어를 x_c라 하면, x_c보다 작은 경우는 불량으로, x_c보다 큰 경우는 정상으로 자료를 예측하여 분류한다. 그러므로 제 I 종 오류율과 제 II 종 오류율은 다음과 같다.

$$\text{제 I 종 오류율} = \sum_{i=1}^{n} I(X_i > x_c)/n, \tag{3.2}$$

$$\text{제 II 종 오류율} = \sum_{j=1}^{m} I(Y_j \leq x_c)/m.$$

분류기준 스코어를 고정함으로써 오분류율을 계산할 수 있으므로 제 I 종 오류율을 5%로 고정하는 분류기준점 x_c를 공통적으로 적용한다. 실제로도 제 I 종 오류의 위험(risk) 관리를 위하여 제 I 종 오류율을 통제한다(예를 들어 5%).

3.2. 모의실험 절차

2절에서 설명한 AR과 AUROC에 대해서 불량률, 표본크기 그리고 제II종 오류율을 적용하기 위한 모의실험 절차는 다음과 같다.

1. 표준정규분포 $N(0,1)$로부터 N개의 난수를 생성하여 N개의 스코어로 간주한다(N = 500, 1,000, 5,000, 10,000).

2. 과정1에서 생성된 스코어를 대상으로 크기 순서대로 나열했을 때 표본불량률 r에 대응하는 값 $x_r = \Phi^{-1}(r)$을 경계로 스코어가 x_r이하인 $n' \approx Nr$자료를 생성하고 n'개 중에서 $n \approx Np$개의 불량을 임의로 추출하여 불량으로 변환한다(p = 0.03, 0.05이고 r = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0).

3. 통계량들을 계산하고 제I종 오류율 = 0.05를 만족하는 "분류기준 스코어" x_c를 경계로 오류표를 작성하고 제II종 오류율을 생성한다.

4. 위 1~3의 과정을 10,000번씩 수행한다.

5. 표본불량률 r을 구분하여 생성된 자료들을 통합하고, 제II종 오류율을 기준으로 평균차이, AUROC, AR 통계량의 평균과 각각의 90과 95 백분위수 값을 산출한다.

자세한 모의실험을 위한 가정과 절차는 박용석과 홍종선 (2008)에 설명되어 있다.

3.3. 모의실험 결과

불량률 p, 표본크기 N, 그리고 제II종 오류율을 기준으로 구한 평균차이(MD), AUROC, AR 통계량들의 결과는 표 2 그리고 표 3과 같다. p = 0.03인 표 2의 결과를 보면 동일한 제II종 오류율에서 표본크기가 커질수록 MD, AUROC, AR 등의 통계량의 평균값은 커진다. 제II종 오류율 10%에 대해서 N = 500인 경우 MD는 1.785, AUROC는 0.960 그리고 AR은 0.921인 반면 N = 10,000인 경우 MD, AUROC 그리고 AR은 각각 1.810, 0.964, 0.928이다.

동일한 표본크기에서 제II종 오류율이 커지면 통계량들은 선형적으로 감소한다. 불량률 p = 0.03에서 AUROC를 토대로 살펴보면 표본크기 N = 1,000인 경우 제II종 오류율이 10%일 때 0.964이고 40%일 때 0.810 그리고 90%일 때 0.548이다. 다른 통계량들도 동일한 패턴을 갖는다.

표본크기와 제II종 오류율을 함께 고려해보면 제II종 오류율이 같고 표본크기가 커지면 MD, AUROC, AR 통계량은 변화가 작아진다. 예를 들어 표 2의 평균차이 결과를 살펴보면 제II종 오류율 30%에서 평균차이 통계량은 N = 500인 경우 1.186, N = 1,000인 경우 1.197 그리고 N = 10,000인 경우 1.195로 표본크기가 커질수록 변화량이 작아짐을 알 수 있다.

MD, AUROC, AR의 90과 95 백분위수의 결과를 정리하면 다음과 같다. 세 통계량의 90과 95 백분위수는 표본크기가 커짐에 따라서 작아지고 작아지는 크기는 평균보다 훨씬 크다. 예를 들어 제II종 오류율 30%에 대한 AUROC의 95 백분위수를 표본크기에 따라서 살펴보면 N = 500에서 0.899, N = 1,000일 때 0.891 그리고 N = 10,000일 때 0.876으로 동일한 제II종 오류율에서 표본크기가 커짐에 따라서 값의 변화가 커진다. 이는 MD, AUROC, AR 통계량들의 판단기준을 고려하는 경우 표본크기를 고려해야 한다는 것을 의미한다.

표 3은 불량률 p = 0.05에 대한 표본크기와 제II종 오류율별 MD, AUROC, AR의 평균 및 90과 95 백분위수 결과이다. 세 가지 통계량들의 평균 및 90과 95 백분위수 결과는 p = 0.03일 때와 유사하기 때문에 설명은 생략하기로 한다.

표 2: 불량률 $p = 0.03$에서 통계량들의 평균 및 90과 95 백분위수

표본 크기	제II종 오류율	의미	MD			AUROC			AR		
			MD_m	$MD_{\beta,90}$	$MD_{\beta,95}$	AUC_m	$AUC_{\beta,90}$	$AUC_{\beta,95}$	AR_m	$AR_{\beta,90}$	$AR_{\beta,95}$
500	10%	Superior	1.785	1.963	2.006	0.960	0.974	0.977	0.921	0.949	0.954
	20%	Excellent	1.429	1.613	1.664	0.909	0.933	0.939	0.818	0.866	0.877
	30%	Very Strong	1.186	1.383	1.440	0.858	0.892	0.900	0.716	0.783	0.799
	40%	Strong	0.985	1.190	1.250	0.806	0.848	0.858	0.612	0.696	0.715
	50%	Very Good	0.809	1.023	1.091	0.753	0.804	0.817	0.506	0.608	0.634
	60%	Good	0.648	0.868	0.936	0.701	0.758	0.774	0.402	0.516	0.548
	70%	Satisfactory	0.594	0.744	0.876	0.674	0.751	0.760	0.348	0.501	0.520
	80%	Marginal	0.343	0.593	0.663	0.594	0.671	0.691	0.189	0.342	0.383
	90%	Poor	0.236	0.482	0.573	0.547	0.636	0.663	0.075	0.272	0.326
	≥ 90%	Doubtful	0.203	0.416	0.504	0.488	0.582	0.606	0.000	0.163	0.212
1,000	10%	Superior	1.809	1.920	1.952	0.964	0.972	0.974	0.928	0.944	0.948
	20%	Excellent	1.444	1.566	1.603	0.913	0.928	0.932	0.825	0.856	0.864
	30%	Very Strong	1.197	1.329	1.369	0.861	0.883	0.889	0.723	0.766	0.779
	40%	Strong	0.997	1.138	1.179	0.810	0.839	0.847	0.620	0.677	0.694
	50%	Very Good	0.820	0.969	1.013	0.757	0.792	0.802	0.515	0.584	0.605
	60%	Good	0.664	0.822	0.867	0.706	0.747	0.759	0.412	0.494	0.519
	70%	Satisfactory	0.511	0.678	0.730	0.654	0.702	0.715	0.308	0.404	0.430
	80%	Marginal	0.356	0.538	0.590	0.601	0.656	0.671	0.203	0.312	0.342
	90%	Poor	0.206	0.392	0.448	0.548	0.610	0.627	0.097	0.221	0.254
	≥ 90%	Doubtful	0.144	0.294	0.352	0.496	0.563	0.582	0.000	0.125	0.164
5,000	10%	Superior	1.810	1.858	1.872	0.964	0.967	0.968	0.928	0.935	0.937
	20%	Excellent	1.444	1.497	1.513	0.912	0.919	0.921	0.825	0.838	0.842
	30%	Very Strong	1.196	1.254	1.271	0.861	0.871	0.874	0.722	0.742	0.748
	40%	Strong	0.996	1.059	1.078	0.810	0.822	0.826	0.619	0.645	0.652
	50%	Very Good	0.823	0.890	0.909	0.758	0.774	0.778	0.516	0.547	0.556
	60%	Good	0.664	0.734	0.754	0.706	0.725	0.730	0.412	0.450	0.460
	70%	Satisfactory	0.512	0.590	0.612	0.655	0.677	0.683	0.310	0.354	0.366
	80%	Marginal	0.360	0.442	0.467	0.603	0.628	0.635	0.206	0.255	0.269
	90%	Poor	0.199	0.288	0.314	0.551	0.578	0.586	0.102	0.156	0.172
	≥ 90%	Doubtful	0.066	0.135	0.163	0.500	0.531	0.538	0.000	0.061	0.076
10,000	10%	Superior	1.810	1.845	1.854	0.964	0.967	0.967	0.928	0.933	0.934
	20%	Excellent	1.444	1.481	1.492	0.912	0.917	0.918	0.825	0.834	0.837
	30%	Very Strong	1.195	1.235	1.247	0.861	0.868	0.870	0.722	0.735	0.739
	40%	Strong	0.996	1.039	1.052	0.809	0.818	0.821	0.618	0.636	0.642
	50%	Very Good	0.822	0.869	0.882	0.758	0.769	0.772	0.515	0.537	0.544
	60%	Good	0.664	0.713	0.728	0.706	0.719	0.723	0.412	0.439	0.446
	70%	Satisfactory	0.513	0.566	0.581	0.655	0.670	0.674	0.310	0.340	0.349
	80%	Marginal	0.361	0.419	0.436	0.603	0.621	0.625	0.206	0.242	0.251
	90%	Poor	0.201	0.265	0.283	0.552	0.571	0.576	0.103	0.142	0.152
	≥ 90%	Doubtful	0.046	0.095	0.115	0.500	0.522	0.528	0.000	0.043	0.056

3.4. 대안적인 판단기준

표 2와 표 3의 판단기준은 불량률, 표본크기 그리고 제II종 오류율을 함께 고려할 수 있으며 상위 90과 95 백분위수를 판단기준으로 가설검정할 수 있다. 가설 검정을 수행하는 방법은 박용석과 홍종선 (2008)에 제시한 K-S 방법과 동일하게 적용할 수 있으므로 본 연구에서는 사례자료를 통한 가설검정 결과만을 고려한다.

표 2와 표 3의 판별력 정도는 총 10개 단계로 판별력이 가장 좋은 Superior부터 가장 판별력이 가장 낮은 Doubtful까지 부여하였다. 표 1은 표본크기와 불량률에 관계없이 모두 동일한 판별력 정도의 의미를 나타내고 있지만 표 2와 표 3에서는 표본크기와 불량률에 따른 판별력 검정기준을 제시하였다. 즉 표 2의 $p = 0.03$인 경우 표본크기 $N = 1{,}000$이고 AUROC가 0.83이면 Very Good에 해당하는 판별력 정도를 나타낸다. 동일한 AUROC 값이 $N = 5{,}000$인 경우로 산출되었다면 Strong으로 해석된다.

표 3: 불량률 $p = 0.05$에서 통계량들의 평균 및 90과 95 백분위수

표본 크기	제II종 오류율	의미	MD			AUROC			AR		
			MD_m	$MD_{\beta,.90}$	$MD_{\beta,.95}$	AUC_m	$AUC_{\beta,.90}$	$AUC_{\beta,.95}$	AR_m	$AR_{\beta,.90}$	$AR_{\beta,.95}$
500	10%	Superior	1.845	1.972	2.010	0.974	0.982	0.984	0.947	0.964	0.968
	20%	Excellent	1.472	1.611	1.652	0.921	0.938	0.943	0.841	0.876	0.885
	30%	Very Strong	1.217	1.367	1.412	0.868	0.893	0.899	0.736	0.785	0.798
	40%	Strong	1.013	1.170	1.217	0.815	0.846	0.854	0.629	0.692	0.709
	50%	Very Good	0.838	1.006	1.052	0.763	0.801	0.812	0.525	0.603	0.624
	60%	Good	0.675	0.850	0.902	0.709	0.755	0.767	0.419	0.510	0.535
	70%	Satisfactory	0.520	0.708	0.760	0.657	0.710	0.724	0.314	0.420	0.448
	80%	Marginal	0.367	0.566	0.627	0.605	0.664	0.682	0.210	0.329	0.363
	90%	Poor	0.220	0.419	0.483	0.551	0.617	0.637	0.102	0.235	0.274
	≥ 90%	Doubtful	0.160	0.330	0.397	0.496	0.571	0.591	0.000	0.141	0.181
1,000	10%	Superior	1.848	1.940	1.965	0.974	0.980	0.981	0.947	0.960	0.963
	20%	Excellent	1.473	1.573	1.602	0.921	0.933	0.936	0.842	0.866	0.872
	30%	Very Strong	1.220	1.327	1.358	0.868	0.886	0.891	0.737	0.772	0.781
	40%	Strong	1.015	1.127	1.158	0.815	0.838	0.845	0.631	0.676	0.689
	50%	Very Good	0.838	0.956	0.991	0.763	0.790	0.798	0.525	0.581	0.597
	60%	Good	0.673	0.796	0.832	0.709	0.741	0.750	0.418	0.483	0.501
	70%	Satisfactory	0.518	0.650	0.690	0.656	0.694	0.704	0.312	0.388	0.409
	80%	Marginal	0.361	0.504	0.542	0.603	0.646	0.657	0.206	0.292	0.315
	90%	Poor	0.202	0.352	0.396	0.550	0.597	0.610	0.100	0.194	0.220
	≥ 90%	Doubtful	0.113	0.232	0.277	0.498	0.550	0.564	0.000	0.101	0.128
5,000	10%	Superior	1.848	1.887	1.898	0.974	0.976	0.977	0.947	0.953	0.954
	20%	Excellent	1.474	1.518	1.530	0.921	0.927	0.928	0.842	0.853	0.856
	30%	Very Strong	1.220	1.267	1.281	0.869	0.876	0.879	0.737	0.752	0.757
	40%	Strong	1.017	1.067	1.081	0.816	0.826	0.829	0.632	0.652	0.658
	50%	Very Good	0.840	0.894	0.909	0.764	0.776	0.779	0.527	0.552	0.559
	60%	Good	0.677	0.734	0.749	0.710	0.725	0.730	0.421	0.451	0.459
	70%	Satisfactory	0.523	0.583	0.601	0.658	0.675	0.680	0.316	0.351	0.360
	80%	Marginal	0.368	0.434	0.453	0.605	0.625	0.631	0.211	0.249	0.261
	90%	Poor	0.205	0.275	0.295	0.553	0.574	0.580	0.105	0.148	0.160
	≥ 90%	Doubtful	0.051	0.107	0.127	0.500	0.524	0.531	0.000	0.047	0.062
10,000	10%	Superior	1.848	1.876	1.884	0.974	0.976	0.976	0.947	0.951	0.952
	20%	Excellent	1.474	1.505	1.514	0.921	0.925	0.926	0.842	0.850	0.852
	30%	Very Strong	1.220	1.253	1.263	0.868	0.874	0.876	0.737	0.748	0.751
	40%	Strong	1.017	1.052	1.062	0.816	0.823	0.825	0.632	0.646	0.650
	50%	Very Good	0.840	0.877	0.888	0.763	0.772	0.774	0.526	0.544	0.549
	60%	Good	0.678	0.718	0.729	0.711	0.721	0.724	0.421	0.442	0.448
	70%	Satisfactory	0.523	0.565	0.577	0.658	0.670	0.673	0.316	0.340	0.347
	80%	Marginal	0.369	0.415	0.427	0.605	0.619	0.623	0.211	0.238	0.246
	90%	Poor	0.206	0.256	0.269	0.553	0.568	0.572	0.105	0.136	0.145
	≥ 90%	Doubtful	0.036	0.075	0.089	0.500	0.517	0.522	0.001	0.035	0.044

또한 불량률 $p = 0.05$이고 표본크기 $N = 1{,}000$에서 K-S 통계량이 0.68이면 Strong으로 해석될 수 있다.

현재 적용되고 있는 표 1을 근거로 한 판별력에 대한 적정성 여부는 Satisfactory에 대응하는 값을 기준으로 한다. 이를 참고로 표 2와 표 3의 표본크기와 불량률에 대해서 Satisfactory보다 크면 판별력이 있다고 판단하고 $p = 0.03$에서 AUROC와 AR의 Satisfactory 기준은 95 백분위수를 기준으로 $N = 500$일 때 각각 0.760, 0.520이고 $N = 5{,}000$일 때 각각 0.683, 0.366이 적용될 수 있다.

실제 사례에 대한 적용을 위해서 홍종선 등 (2008)에 수록되어 있는 사례의 통계량을 적용해보면 사례자료는 1994년부터 2005년까지 대기업 대상 중 매출액 1,000억 이상의 대기업에 관한 연도별로 4,268건(정상: 4,101, 부도: 167)의 재무자료를 바탕으로 생성한 신용평가 모형으로 판별력에 대한 검정 통계량들이 MD = 1.404, AUROC = 0.891, AR = 0.782의 값을 갖는 모형이다. 불량률은 모형생성을 위해 수집된 전체자료 중에서 불량대상의 비율로써 $\hat{p} \approx 0.03$이다. 그리고 전체 표본크기는 4,268건이

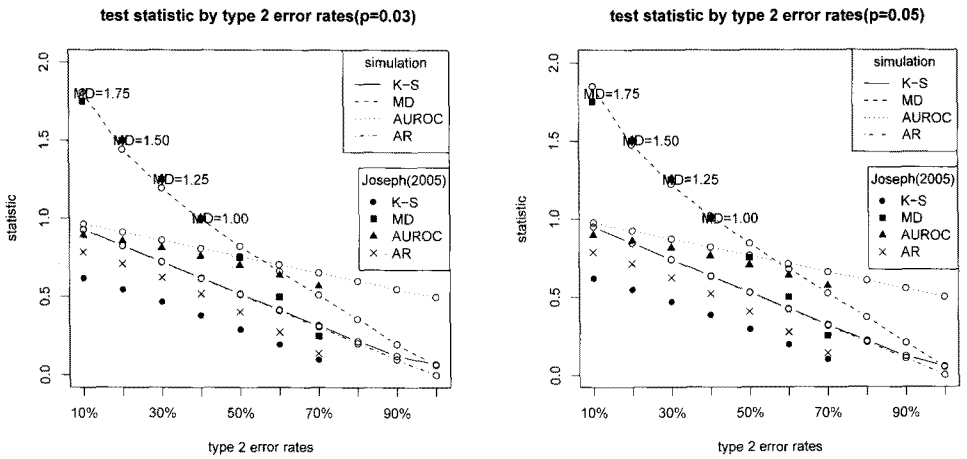

그림 1: $p = 0.03$과 $p = 0.05$에 대한 제II종오류율별 판단기준 비교

므로 N = 5,000을 적용한다. 불량률 p = 0.03이므로 표 2를 고려하여 N = 5,000에서 MD, AUROC, AR 값들을 비교해보면 세 가지 통계량들의 판단기준은 다음과 같다.

$$MD_{20\%,0.95} = 1.513 > 1.404 > MD_{30\%,0.95} = 1.271,$$
$$AUC_{20\%,0.95} = 0.921 > 0.891 > AUC_{30\%,0.95} = 0.874,$$
$$AR_{20\%,0.95} = 0.842 > 0.782 > AR_{30\%,0.95} = 0.748.$$

유의수준을 5%로 하여 귀무가설을 기각하는 임계기준들은 $MD_{30\%,0.95}$ = 1.271, $AUC_{30\%,0.95}$ = 0.874 그리고 $AR_{30\%,0.95}$ = 0.748이다. 이 값들은 모두 제II종 오류율 30% 수준의 값들로 제II종 오류율을 30% 수준에서 판별력이 있는 모형이라고 결론 내릴 수 있다. 판별력에 대한 의미를 기준으로 살펴보면 MD, AUROC, AR 세 개의 통계량은 모두 'Very Strong'에 해당하는 값으로 모형의 판별력이 매우 높다고 해석할 수 있다.

3.5. 적합성 검정기준 통계량들의 비교

네 가지 적합성 기준의 비교를 위해서는 표 1과 2 그리고 표 3에 모두 존재하는 동일한 기준이 필요하다. 그런데 표 1은 MD를 기준으로 산출되었고 표 2는 제II종 오류율을 기준으로 산출되어있기 때문에 직접 비교하기 어렵다. 하지만 표 2의 결과를 보면 MD의 값 1.75, 1.50, 1.25, 1.00은 각각 제II종 오류율 10%, 20%, 30%, 40%일때의 MD 값과 유사하다. 각 제II종 오류율에 대응하는 MD, AUROC, AR 그리고 K-S 통계량을 나타내면 그림 1과 같다. 그림 1의 왼쪽 그림은 불량률 p = 0.03인 경우이고 오른쪽 그림은 p = 0.05인 경우이다.

그림 1의 왼쪽 그림인 p = 0.03의 결과를 살펴보면 MD, AUROC, AR, K-S 통계량에 대한 표 1의 기준은 점으로 표시하였고 표 2의 결과는 선으로 연결하여 나타내었다. MD, AUROC, AR, K-S 등 네 개의 통계량은 표 1과 2의 결과 모두 유사한 패턴을 갖는다. 제II종 오류율이 감소함에 따라서 일정한 감소량 즉 유사한 기울기를 갖는다. 다만 MD는 표 1에서 완전히 선형적으로 감소하는 현상이 나타내는데 본 연구에서 실험된 결과인 표 2은 제II종 오류율이 10%와 20%로 작은 경우는 감소량이 크다가 제II종 오류율이 60%로 가까워질수록 감소량이 작아진다. Joseph (2005)가 제안한 표 1의 기준이 모

의실험 결과인 표 2보다 약간 낮다. 특히 K-S 통계량은 다른 통계량들보다 표 1의 기준이 낮으며 K-S 통계량과 AR 통계량은 판단기준이 거의 일치한다. 그러므로 표본크기의 차이가 존재하는 경우 표 1의 판단기준은 K-S 통계량보다 AR과 AUROC의 판단기준이 더 안정적이라고 할 수 있다.

4. 결론

본 연구에서는 모의실험 결과를 통해서 Joseph (2005)가 제안한 표 1의 일반적인 판단기준에 대한 대안으로 표본크기 N, 불량률 p 그리고 제II종 오류율에 근거한 대안적인 판단기준 90과 95 백분위수를 제시하였다. 대안적인 판단기준은 상위 90과 95 백분위수를 기준으로 허용할 수 있는 제II종 오류율 범위 내에서 가설 검정할 수 있다. 이와 함께 표 1의 의미를 참고한 10단계의 의미를 부여하고 표본크기와 불량률에 따라서 다르게 해석할 수 있는 통계량의 판별력 정도에 대한 의미를 제시하였다. K-S, AUROC, AR 그리고 MD 통계량들의 관계를 그림 1을 통해서 살펴보았다. 네 가지 판별력 판단기준 통계량들 중에서 K-S 통계량의 판단기준이 타 통계량들보다 차이가 크고 모의실험 결과 표 1에서의 판단기준과 다르게 K-S와 AR 통계량은 거의 같은 값을 갖는다. 이는 불량과 정상의 표본크기의 차이가 존재하는 경우 표 1과 2 그리고 3의 판단기준 차이를 고려할 때 AR이 K-S의 차이보다 작기 때문에 더 안정된 판단기준이라고 판단할 수 있다. 또한 두 개 이상의 신용평가모형들에 대한 상대적인 판단기준으로 적용되었던 AUROC와 AR의 판단기준은 표 2와 3을 통해 판별력 검정에 대한 절대적인 판단기준으로 적용할 수 있다.

참고 문헌

박용석, 홍종선 (2008). 신용평가모형에서 콜모고로프-스미르노프 검정기준의 문제점, <한국통계학회 논문집>, **15**, 1013–1026.

홍종선, 이창혁, 김지훈 (2008). 범주형 재무자료에 대한 신용평가모형 검증 비교, <한국통계학회 논문집>, **15**, 615–631.

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

Engelmann, B., Hayden, E. and Tasche, D. (2003). *Measuring the Discriminative Power of Rating Systems*, Discussion paper Series 2: Banking and Financial Supervision.

Faraggi, D. and Reiser, B. (2002). Estimation of the area under the ROC curve, *Statistics in Medicine*, **21**, 3093–3106.

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.

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

Sobehart, J. R., Keenan, S. C. and Stein, R. M. (2000). Benchmarking quantitative default risk models: A validation methodology, *Moody's Investors Service*.

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

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

Tasche, D. (2006). Validation of internal rating systems and PD estimates, *eprint arXiv:physics/0606071*.

Wilkie, A. D. (2004). *Measures for Comparing Scoring Systems, In Readings in Credit Scoring-Recent Developments, Advances, and Aims*, Oxford Finance.

2008년 11월 접수; 2009년 1월 채택

Criterion of Test Statistics for Validation in Credit Rating Model

Yong Seok Park[a], Chong Sun Hong[1,b], Han Seung Lim[c]

[a]Research Institute of Applied Statistics, Sungkyunkwan Univ.,
[b]Dept. of Statistics, Sungkyunkwan Univ., [c]RMS 2 Team, Korea Ratings

Abstract

This paper presents Kolmogorov-Smirnov, mean difference, AUROC and AR, four well known statistics that have been widely used for evaluating the discriminatory power of credit rating models. Criteria for these statistics are determined by the value of mean difference under the assumption of normality and equal standard deviation. Alternative criteria are proposed through the simulations according to various sample sizes, type II error rates, and the ratio of bads, also we suggest the meaning of statistic on the basis of discriminatory power. Finally we make a comparative study of the currently used guidelines and simulated results.

Keywords: Accuracy ratio, area under ROC, discriminatory power, mean difference, type II error rates.

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