Communications of the Korean Statistical Society Vol. 15, No. 6, 2008, pp. 1013–1026

신용평가모형에서 콜모고로프-스미르노프 검정기준의 문제점

박용석[1], 홍종선[2]

요 약

신용평가모형의 판별력에 대한 적합성 검정방법으로 콜모고로프-스미르노프(K-S) 통계량이 널리 사용되고 있다. K-S 통계량을 통한 모형의 판별력 판단기준으로는 표본수에 의존하는 K-S 검정통계량의 임계값보다 매우 큰 기준인 0.3~0.4의 수준이 일반적으로 적용된다. 본 논문에서는 모의실험을 통해 일반적 판단기준의 타당성을 살펴보았다. 모의실험 결과 국내에서 개발된 대부분의 신용평가모형의 결과를 바탕으로 구한 K-S 통계량은 현재 적용하고 있는 판단기준보다 큰 값을 갖는다는 것을 발견하였다. 따라서 어떠한 신용평가모형이라도 좋은 판별력을 갖는다고 해석할 수 있다. 본 연구에서는 표본크기와 불량률 그리고 제II종오류율에 따른 대안적인 임계값을 제안한다.

주요용어: 신용평가모형; 임계값; 콜모고로프-스미르노프 통계량; 타당성; 판별력.

1. 서론

신용평가모형(credit rating model)의 타당성검증(validation)은 모형의 설계와 방법론에 해당하는 질적인 부분과 모형의 성능(performance)에 해당하는 양적인 부분으로 나눌 수 있으며, 이는 다시 판별력(discriminatory power)과 안정성, 등급의 계량화 즉, 부도확률(probability of default) 추정에 대한 성능으로 나눌 수 있다. 특히 신용평가모형의 판별력 기준으로 널리 쓰이고 있는 통계적인 기준으로 χ^2 검정통계량, 콜모고로프-스미르노프(Kolmogorov-Smirnov: K-S) 검정통계량, 평균차이(mean difference) 검정통계량, 부도 50%에 대한 정상의 누적비율($1-\mathrm{PH}$), AR(accuracy ratio), Gini 계수, ROC(receiver operating characteristic), 정보량(information), K-L(Kullback-Leibler) 통계량 등 여러가지 방법들이 사용되고 있다.

Wilkie (1992)와 Hand (1994)는 신용평가모형의 판별력 검정 통계량으로써 T 통계량과 정보량에 대해서 연구하였다. 이들 연구와는 다른 관점에서 예측과 실제의 불량과 정상에 대한 2차원 분할표(contingency table)를 이용한 시각화 방법인 CAP과 ROC 곡선 그리고 그에 대응하는 AR과 AUROC의 특성 및 관계에 대한 연구들로 Engelmann 등 (2003a, 2003b), Fernandes (2005) 등이 있다. Thomas 등 (2002)은 K-S 통계량과 그 외 추가적인 판별력검정 방법들에 대해서 연구하였고 Wilkie (2004)는 정규분포 하에서 평균

1) (110-745) 서울 종로구 명륜동 3가 53 성균관대학교 응용통계연구소, 연구원.
2) (110-745) 서울 종로구 명륜동 3가 53 성균관대학교선 통계학과, 교수. 교신저자: cshong@skku.ac.kr

표 1.1: 정규분포 가정에 기초한 K-S 통계량의 판단기준

의미	평균차이(MD)	K-S
Random	0.00	0.00
Doubtfull	0.25	0.10
Poor	0.50	0.20
Marginal	0.75	0.29
Satisfactory	1.00	0.38
Good	1.25	0.47
Very Good	1.50	0.55
Strong	1.75	0.62
Very Strong	2.00	0.68
Excellent	2.25	0.74
Excellent	2.50	0.79
Excellent	2.75	0.83
Superior	3.00	0.87

차이 검정통계량, 부도 50%에 대한 정상의 누적비율, K-S 검정통계량, Gini 계수, 정보량에 대한 검증 기준을 생성하였다. Joseph (2005)는 Wilkie (2004)의 결과에 AR, ROC, K-L 통계량들을 추가하여 보다 확장된 모형 판별력 검정 기준을 표 1.1과 같이 제안하였다. Joseph (2005)는 이들 통계량들에 대해서 불량과 정상이 정규분포라는 가정 하에서 평균차이를 기준으로 한 13구간의 판단 기준을 설정하고 그에 따른 K-S 통계량의 적정성 기준으로 0.3 또는 0.4 이상이 되어야 함을 설명하고 있다.

비모수(nonparametric) 검정방법인 K-S 검정방법은 임계값(critical value)을 기준으로 가설을 검정하며, 임계값은 표본의 크기에 따라 달라진다. 소표본인 경우에는 표본수에 따른 임계값을 적용하고, 대표본인 경우에는 극한분포(limiting distribution)를 이용한 근사 임계값을 이용한다. 표본의 크기 n과 m이 충분히 큰 두 표본에 대해서 유의수준 5%에서 K-S 통계량의 임계값은 $1.22\sqrt{(n+m)/(nm)}$을 이용한다 (Daniel 1990; 송문섭 등, 2003 참고). 임계값을 기준으로 산출된 K-S 통계량이 임계값보다 크면, 두 모집단의 분포함수가 동일하다는 귀무가설 $H_0 : F_1(x) = F_2(x)$을 기각한다. 이와 같이 K-S 통계량의 임계값은 표본의 크기에 의존하고 두 표본의 경우 두 표본크기의 차이에 의존하게 된다. 표본크기가 커지면 K-S 통계량의 임계값은 매우 작아지게 된다. 이는 모형 판별력에 대한 검정이 민감해지기 때문에 동일한 두 분포함수의 귀무가설을 쉽게 기각하게 하는 문제를 야기한다.

실제 신용평가모형 구축시 사용되는 자료는 일반적으로 표본크기가 매우 크다. 기업을 대상으로 하는 자료는 물론이고 개인을 대상으로 하는 자료는 표본크기가 더 크다. 따라서 신용평가모형에 적용하게 되는 K-S 검정은 그 임계값이 매우 작아지게 된다. 예를 들면 전체 10,000개의 표본에 대해서 정상과 불량 이 각각 9,500, 500이라면 유의수준 5%에서 K-S 검정의 임계값은 0.0624로 매우 작은 값이다. 그러나 신용평가모형에서 K-S 통계량 값은 일반적으로 임계값보다 매우 큰 값을 나타낸다. 즉 두 분포함수가 동일하다는 귀무가설을 대부분 기각하게 된다. 이러한 문제로 인해 실제 신용평가모형의 판별력을 평가할 때

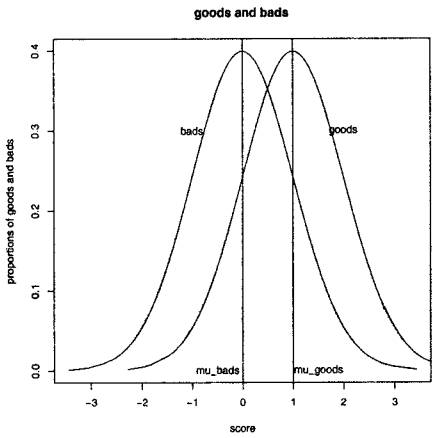

그림 1.1: 정상과 불량의 비율

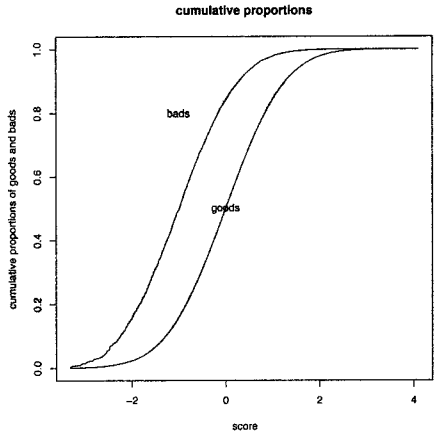

그림 1.2: 정상과 불량의 누적비율

K-S 통계량의 판단기준으로는 Joseph (2005)의 기준을 이용하고 있다. Joseph (2005)가 제안하고 있는 모형 판별력 판단기준은 표 1.1과 같다.

표 1.1에 제시되어있는 K-S 통계량의 판단기준을 산출하는 방법에 대해서 살펴보자. 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 각각 σ_g, σ_b이다. 여기서 표준편차는 동일하다고 가정하여 σ라 하면, 표 1.1의 평균차이(MD)는 다음과 같이 나타낼 수 있다(자세한 내용은 Wilkie (2004) 참조).

$$\mathrm{MD} = \frac{\mu_g - \mu_b}{\sigma}. \tag{1.1}$$

그러므로 식 (1.1)의 평균차이를 이용하여 식 (1.2)에 의해 K-S 통계량을 구할 수 있으며 그 결과가 표 1.1에서 적용되는 판단기준이다.

$$\begin{aligned} \text{K-S} &= \Phi\left(\frac{\mathrm{MD}}{2}\right) - \Phi\left(-\frac{\mathrm{MD}}{2}\right) \\ &= 2\Phi\left(\frac{\mathrm{MD}}{2}\right) - 1. \end{aligned} \tag{1.2}$$

위 내용을 그림을 이용해서 나타내면 그림 1.1과 1.2와 같다.

표 1.1과 같이 현재 이용되고 있는 일반적인 K-S 통계량의 판단기준은 '정상'과 '불량'이 정규분포를 따른다는 가정과 두 분포의 표준편차가 동일하다는 가정 하에서 산출된 판단기준이다. 이러한 판단기준을 토대로 현재 이루어지고 있는 K-S 통계량의 판별력 판단 방법을 살펴보면 약 0.38이 넘어야 'Satisfactory' 수준이고 'Good'의 경계값은 0.47의 값을 나타내고 있다. 0.55 이상의 값을 가지면 'Very Good'이라 평가하고 0.62보다 크면 'Strong'이라 판정한다. 이를 요약하여 일반적으로 사용하는 평가 기준은 다음과 같다: K-S 통계량값이 0.2 이하이면 판별력이 낮은 모형으로 판단을 하고, 0.2~0.4는 적정하다고

판단한다. 그리고 0.4~0.5의 값을 나타내면 모형의 판별력이 좋다고 판단하며 0.5 이상이면 판별력이 매우 좋다고 결론을 내린다.

Joseph (2005)가 제시한 표 1.1의 판단기준은 두 분포가 정규분포이고 표준편차가 동일하다는 가정 하에서 평균차이를 기준으로 설정된 판단기준이다. 그런데 일반적으로 신용평가모형 수립을 위해 수집된 자료에서 불량은 과거 심사에서 정상으로 판정된 대상에서 발생하였기 때문에 그 비율이 3% 또는 5%일 정도로 매우 작은 경우가 일반적이다. 이렇게 '정상'과 '불량'의 표본크기의 차이가 커지면 두 집단의 표준편차가 같다는 가정을 만족하기가 어려워진다. 그러므로 실제 평가모형에서는 표본수를 고려하여야 한다. 하지만 표 1.1을 기준으로 판단하고 있는 방법은 현실적으로 불량과 정상의 표본 수의 차이와 전체 표본크기를 고려하지 않는 문제점을 갖고 있다.

본 연구에서는 실제 신용평가모형 자료와 유사한 상황과 조건을 부여한 자료를 생성하고 모의실험을 통해 표 1.1의 K-S 통계량의 판단기준에 대한 문제점들을 파악하고 개선점을 제안하고자 한다. 이를 위해서 정규분포 하에서 난수를 생성하여 스코어로 간주하고, 그 중에서 정상과 불량의 비율을 현실적이 되도록 95대 5와 97대 3으로 설정한다. 표본크기 N과 불량률 p를 기준으로 K-S 통계량의 분포를 살펴보고 제Ⅱ종오류율을 고려한 표 1.1의 평가기준에 대한 대안적인 임계값(alternative critical value)을 제시한다.

본 논문의 구성은 다음과 같다. 2절에서는 연구방법에 대해서 설명한다. 실제 신용평가 자료와 동일한 자료를 생성하기 위한 난수의 생성방법을 포함한 모의실험 설계와 모의실험 절차에 대해서 설명한다. 3절에서는 모의실험 결과를 통해 현재 적용되는 K-S 통계량의 분포를 살펴보고 오류와 오류율을 구한다. 그리고 K-S 통계량의 판단기준과 비교해보고 대안이 되는 판단기준을 제시하며 사례를 통한 대안적 판단기준의 적용방법을 살펴본다. 4절에서는 모의실험 결과에 대해서 정리하고 토론한다.

2. 연구방법

2.1. 모의실험 설계

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 전체 불량률(total probability of bads)은 $p = P[Z = 1] = 1 - P[Z = 0]$로 정의되며, 전체자료를 N개라 하면 불량의 개수는 $n \approx Np$이고 정상의 개수는 $m = N - n$으로 표현된다.

전체 스코어 분포함수(distribution function)를 $F(x) = P[X \leq x]$로 정의하자. 그러면 불량의 스코어 분포함수와 정상의 스코어 분포함수는 각각 $F_b(x) = P[X \leq x \mid Z = 1]$와 $F_g(x) = P[X \leq x \mid Z = 0]$으로 정의되고 다음과 같은 불량과 정상의 분포함수 형태로 분할된다 (Tasche, 2006).

$$F(x) = pF_b(x) + (1 - p)F_g(x), \tag{2.1}$$

여기서 p는 전체 불량률이고 $p \in (0, 1)$이다.

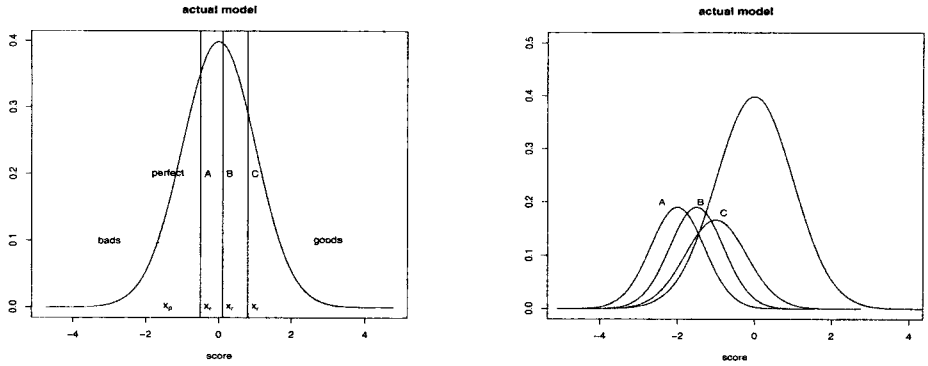

그림 2.1: 모형의 판별력과 r의 관계

신용평가모형이 불량과 정상을 정확하게 구분하는 완전한 모형(perfect model)이라면 n개의 불량에 대한 스코어가 정상의 스코어보다 모두 작은 값을 갖게 된다. 이러한 경우 불량과 정상의 스코어를 각각 $\{X_1, \ldots, X_n\}$, $\{Y_1, \ldots, Y_m\}$이라 하면 두 집단에 대한 스코어는 $x_{(n)} \leq y_{(1)}$의 관계를 나타낸다. 그러므로 모형의 판별력이 완전하지 않으면 불량과 정상 자료는 혼합되게 된다. 혼합되는 정도는 다음의 식 (2.2)를 만족하는 r을 통해 나타낼 수 있다.

$$F\left(x_{(n)}\right) = r, \tag{2.2}$$

여기서 r은 실제모형(actual model)에서 전체 자료 중 $x_{(n)}$ 이하의 값을 갖는 자료의 비율로 표본불량률이라고 한다. 그리고 $F_b(x_{(n)}) = 1$이다. 식 (2.2)에서 만약 $x_{(n)} = F^{-1}(r) \leq y_{(1)}$이면 $r = p$인 경우이고, $x_{(n)} > y_{(1)}$이면 $r > p$이며 정상과 불량이 혼합된 경우를 의미한다. 표본불량률 r은 다음과 같이 정의된다.

$$r = \frac{\sum_{i=1}^{n} I(X_i \leq x_{(n)}) + \sum_{j=1}^{m} I(Y_j \leq x_{(n)})}{N} = \frac{n + \sum_{j=1}^{m} I(Y_j \leq x_{(n)})}{N}. \tag{2.3}$$

불량률 p와 표본불량률 r의 관계에서 p과 r의 차이가 작으면 모형의 판별력이 좋다는 것을 의미다. 하지만 p과 r의 차이가 크더라도 $n' \approx Nr (n' > n)$개 중에서 선택된 n개의 불량 발생 위치에 따라서 판별력은 달라진다. 이는 불량 발생 위치에 따라서 다른 모형들로 고려할 수 있고 K-S 통계량이 발생할 수 있는 범위가 넓다는 것을 의미한다. 본 연구에서는 식 (2.2)와 (2.3)을 만족하는 표본불량률 r을 조정함으로써 불량과 정상의 혼합정도를 달리하는 모의실험 자료를 생성한다.

그림 2.1은 r을 이용하여 불량과 정상이 혼합된 정도가 다른 자료를 생성하는 방법을 나타낸 것이다. 왼쪽 그림에서 점선은 완전한 모형에서의 불량을 구분하는 경계점을 의미하고 실선은 실제 모형에서 r에 의한 경계점을 나타낸다. 모형 A의 r은 모형 B의 r보다 작고, 모형 B의 r은 모형 C의 r보다 작은 경우의 모형을 표시하였다. 그림 2.1의 오른쪽 그림은 왼쪽 그림에서 r에 의해 설정된 판별력이 다른 모형 A, B, C의 불량과 정상이 혼

표 2.1: 혼동행렬

		예측		
		불량	정상	
실제	불량	정분류	제 I 종오류	n
	정상	제 II 종오류	정분류	m

합되는 정도를 표시하였다. 그림 2.1과 같은 형태로 r을 통해 판별력이 다른 모형이 생성된다.

모형에 대한 판별력 검정의 다른 방법으로 오류율(misclassification rates)을 고려할 수 있다. 오류율은 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. 예측된 불량과 정상 그리고 실제 불량과 정상 사이의 오류표(misclassification table) 또는 혼동행렬(confusion matrix)을 작성하면 분류기준 스코어에 의해 표 2.1과 같이 정의된다.

표 2.1에서 보는 것처럼 분류기준 스코어에 의한 오류는 제 I 종오류와 제 II 종오류로 고려할 수 있고 제 I 종오류율과 제 II 종오류율은 다음과 같은 식으로 나타낸다.

$$\text{제 I 종오류율} = P[\text{정상으로예측} \mid \text{실제불량}] = \frac{\sum_{i=1}^{n} I(X_i > x_c)}{n},$$

$$\text{제 II 종오류율} = P[\text{불량으로예측} \mid \text{실제정상}] = \frac{\sum_{j=1}^{m} I(Y_j \le x_c)}{m}. \tag{2.4}$$

그러므로 전체 오류율은 다음과 같다.

$$\text{전체오류율} = \frac{\sum_{i=1}^{n} I(X_i > x_c) + \sum_{j=1}^{m} I(Y_j \le x_c)}{N}.$$

오류율은 분류기준 스코어 x_c에 따라서 달라진다. 따라서 오류율을 평가하기 위해서는 기준을 적용하여 고정시켜 비교할 필요가 있다. 일반적으로 분류기준 스코어는 다음과 같은 기대비용함수를 최소로 하는 지점을 경계로 설정된다 (Koh, 1992).

$$\text{기대비용} = p \times C_1 \times \text{제 I 종오류} + (1-p) \times C_2 \times \text{제 II 종오류}, \tag{2.5}$$

여기서 C_1과 C_2는 각각 제 I 종오류와 제 II 종오류가 발생함으로써 생기는 비용을 말한다. 식 (2.5)의 기대비용함수를 최소로 하는 것은 불량률 p와 제 I 종오류와 제 II 종오류의 비용을 모두 고려해야한다는 것을 의미한다.

신용평가모형에서 제 I 종오류율의 손실(loss)은 대출 또는 금융거래를 허용하였을 때 거래자가 이를 상환하지 못함으로써 발생하는 손실이다. 제 II 종오류율로 인해 발생하는

손실은 정상인 거래자 혹은 기업의 대출 또는 금융거래를 거절해서 발생하는 잠재적인 이자 등에 대한 손실 등이다. 실제 신용평가모형에서는 제 I 종오류율의 위험(risk) 또는 손실이 제 II 종오류율보다 크기 때문에($C_1 > C_2$) 일반적으로 제 I 종오류율을 고정하는 분류기준 스코어를 설정한다. 본 연구에서는 제 I 종오류율 = 0.05로 고정하는 x_c를 분류기준 스코어로 적용하였다. 그러므로 식 (2.5)에서 불량률 p, 제 I 종오류율, C_1 그리고 C_2는 고정되어있는 값이기 때문에 제 II 종오류율에 의해서 전체 기대비용이 결정된다. 따라서 신용평가모형의 사용자는 이 제 II 종오류율을 고려함으로써 전체기대비용을 산출하고 손실여부를 판단할 수 있다. 다른 정보들이 고정된 상태에서 C_2가 크지 않다면 제 II 종오류율이 크더라도 신용평가모형의 사용자는 모형을 수용할 수 있을 것이다.

이제 불량과 정상의 혼합된 정도를 조정하는 표본불량률 r과 제 II 종오류율 그리고 전체오류율의 관계를 살펴보자. 먼저 여기서 제 I 종오류율은 0.05로 고정되었으므로 r과 제 II 종오류와의 관계를 살펴보자. 분류기준 스코어와 불량 중에서 가장 스코어가 큰 값의 관계는 $x_c \leq x_{(n)}$이다. 그리고 앞에서 정의한 것처럼 전체 자료를 N, 불량과 정상을 각각 $n \approx Np$, $m = N - n$이라 하면 식 (2.2)와 (2.3)에 의해서 식 (2.6)과 같이 정리될 수 있다.

$$\frac{N}{m} \times r - \frac{n}{m} \geq \text{제 II 종오류율}, \tag{2.6}$$

여기서 분류기준 스코어 x_c와 $x_{(n)}$은 $x_c \leq x_{(n)}$이지만 일반적으로 $x_c \approx x_{(n)}$인 분류기준스코어가 적용될 때 r과 제 II 종오류율은 근사적으로 선형적인 관계가 존재한다. 전체오류율과 제 II 종오류율은 다음과 같은 관계가 존재한다.

$$\begin{aligned} \text{제 II 종오류율} &= \frac{N}{m} \times \text{전체오류율} - \frac{n}{m} \times \text{제 I 종오류율} \\ &= \frac{N}{m} \times \text{전체오류율} - \frac{n}{m} \times 0.05. \end{aligned} \tag{2.7}$$

식 (2.7)에서 불량과 정상 자료수의 비인 n/m은 3/97또는 5/95이므로 매우 작은 값을 갖고 전체오류율과 제 II 종오류율은 선형적인 관계를 갖으며 r과 전체오류율은 선형적인 관계가 존재한다. 따라서 r을 조정하는 것은 제 II 종오류율을 조정하는 것과 동일하므로 r을 조정함으로써 다른 오류율을 생성할 수 있다.

2.2. 모의실험절차

모의실험은 실제 현장에서 이용되고 있는 스코어의 형태를 그대로 반영하기 위해 실제환경과 유사하게 설계한다. 임의의 정상과 불량의 자료를 현실적인 자료의 형태로 생성하기 위해서 다음 절차를 따른다.

1. 표준정규분포 $N(0,1)$로부터 N개의 난수를 생성하여 N개의 스코어로 간주한다(N = 500, 1,000, 5,000, 10,000, 50,000).

   신용평가모형 구축에 사용되는 자료는 기업체의 경우 대기업은 약 5,000개 정도의 표본이 이용되고 그 밖의 소규모 업체나 개인에 대한 평가모형은 더 많은 자료를 이용한다. 그러므로 이러한 표본크기를 대표하기 위해서 500에서부터 50,000까지의 표본크기를 설정하였다.

2. 과정1에서 생성된 스코어를 대상으로 크기 순서대로 나열했을 때 식 (2.2)를 만족하는 표본불량률 r에 대응하는 자료를 $x_r = \Phi^{-1}(r)$을 경계로 스코어가 x_r 이하인 $n' \approx Nr$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).

   N=1,000이고 $p = 0.03$에 대해서 $r = 0.1$인 경우 예를 들어보자. 불량의 수는 $n \approx Np = 1,000 \times 0.03 = 30$이다. $r = 0.1$이므로 1,000개의 자료 중 스코어를 기준으로 10%에 해당하므로 $\Phi^{-1}(0.1) = -1.28$ 이하의 스코어 값을 갖는 자료에서 불량을 추출한다. 즉 $n' \approx Nr = 1,000 \times 0.1 = 100$개가 되고 이 $n' \approx 100$개의 자료 중에서 30개 자료를 임의로 선택하여 불량으로 지정한다. 즉 n'에 선택된 자료는 30개의 불량과 70개의 정상이 혼합되어 있고 -1.28보다 큰 $N - n' \approx 900$개의 자료는 모두 정상으로 생성되어 총 1,000개의 자료가 생성된다. 실제 현장에서는 스코어가 아주 큰 경우에도 불량이 발생할 수 있다. 하지만 이는 특수한 경우로 분석가들에 의해 여러 가지 방법들로 통제하게 된다. 그러므로 r이 0.7 이상에서 불량이 발생하는 것은 일반적인 상황이 아니라고 고려하여 본 연구에서는 r을 0.1에서 0.7까지 적용하여 자료를 생성하였다.

3. K-S 통계량을 계산하고 제 I 종오류율 = 0.05를 만족하는 "분류기준 스코어" x_c를 경계로 오류표를 작성하고 제 II 종오류율을 생성한다.

4. 위 1~3의 과정을 10,000번씩 수행한다.

5. 표본불량률 r을 구분하여 생성된 자료들을 통합하고, 제 II 종오류율 10%, 20%, 30%, 40%, 50%, 60%를 기준으로 K-S 통계량 평균과 상위 10%와 5% 값인 u_{90}과 u_{95}를 산출한다.

   과정2에서 표본불량률 r의 통제는 판별력이 다른 자료들을 생성하는데 목적이 있으며 식 (2.6)에 의해서 동일한 r에 대해서도 다른 제 II 종오류율을 나타낼 수 있기때문에, 즉 r이 크더라도 제 II 종오류율은 작을 수 있으므로 판별력 정도를 제 II 종오류율을 기준으로 살펴본다. 그러므로 다양한 r들에 의해 생성된 자료를 통합하고 제 II 종오류율을 기준으로 구분하여 판단기준을 산출한다.

3. 모의실험 결과

3.1. K-S 통계량

2절에서 언급한 모의실험 절차를 이용하여 오분류된 자료를 생성한 후 불량률 p, 표본크기 N 그리고 제 II 종오류율을 기준으로 구한 K-S 통계량 결과는 표 3.1과 같다.

일반적인 K-S 통계량은 오류율의 정보를 반영하지 못하지만 모의실험된 결과 표 3.1은 표본크기와 제 II 종오류율의 정보를 기준으로한 K-S 통계량 값을 나타낸다. 표 3.1은 불량률 p를 각각 0.03과 0.05에 대해서 고려한 결과이다. 표 3.1의 K-S 통계량은 p가 0.03과 0.05일 때 모두 제 II 종오류율이 커짐에 따라 K-S 통계량은 감소한다. 그리고 표본크기가

표 3.1: 제II종오류율별 K-S 통계량 평균 및 90과 95백분위수

표본크기	제II종오류율	$p=0.03$			$p=0.05$		
		K-S	백분위수		K-S	백분위수	
			u_{90}	u_{95}		u_{90}	u_{95}
500	10%	0.9268	0.9476	0.9520	0.9492	0.9623	0.9661
	20%	0.8322	0.8616	0.8701	0.8476	0.8705	0.8763
	30%	0.7364	0.7753	0.7869	0.7462	0.7749	0.7829
	40%	0.6407	0.6843	0.6934	0.6455	0.6786	0.6875
	50%	0.5451	0.5905	0.5971	0.5461	0.5830	0.5924
	60%	0.4526	0.4969	0.5150	0.4489	0.4895	0.4979
1,000	10%	0.9302	0.9411	0.9442	0.9485	0.9579	0.9602
	20%	0.8309	0.8477	0.8528	0.8452	0.8607	0.8654
	30%	0.7319	0.7536	0.7606	0.7420	0.7616	0.7672
	40%	0.6335	0.6601	0.6721	0.6384	0.6614	0.6674
	50%	0.5353	0.5672	0.5814	0.5358	0.5608	0.5677
	60%	0.4382	0.4765	0.4903	0.4342	0.4613	0.4696
5,000	10%	0.9284	0.9332	0.9344	0.9476	0.9516	0.9527
	20%	0.8259	0.8329	0.8349	0.8427	0.8495	0.8515
	30%	0.7235	0.7319	0.7344	0.7379	0.7462	0.7483
	40%	0.6214	0.6309	0.6339	0.6332	0.6423	0.6449
	50%	0.5195	0.5298	0.5331	0.5287	0.5383	0.5411
	60%	0.4179	0.4290	0.4330	0.4243	0.4344	0.4373
10,000	10%	0.9281	0.9314	0.9323	0.9475	0.9503	0.9512
	20%	0.8253	0.9303	0.8318	0.8424	0.8472	0.8484
	30%	0.7226	0.7286	0.7302	0.7374	0.7433	0.7449
	40%	0.6201	0.6266	0.6284	0.6323	0.6386	0.6403
	50%	0.5175	0.5246	0.5266	0.5275	0.5342	0.5362
	60%	0.4152	0.4224	0.4247	0.4227	0.4294	0.4316
50,000	10%	0.9279	0.9294	0.9298	0.9474	0.9487	0.9491
	20%	0.8249	0.8270	0.8277	0.8422	0.8443	0.8449
	30%	0.7218	0.7244	0.7251	0.7370	0.7396	0.7403
	40%	0.6189	0.6217	0.6225	0.6317	0.6345	0.6353
	50%	0.5159	0.5188	0.5196	0.5266	0.5295	0.5305
	60%	0.4130	0.4160	0.4169	0.4214	0.4244	0.4252

커짐에 따라 $p=0.03$이 $p=0.05$보다 작은 값을 갖는다. 예를 들어 표본크기 $N=500$이고 $p=0.03$인 경우 K-S 통계량이 제II종오류율 60%까지 고려했을 때 0.4526이고 $p=0.05$인 경우는 0.4489이다. 표본크기 $N=50{,}000$이고 $p=0.03$인 경우는 0.4130이고 $p=0.05$인 경우는 0.4214이다.

제II종오류율이 작을 때에 K-S 통계량은 표본크기에 영향을 덜 받는다. 표 3.1에서 $p=0.03$이고 제II종오류율이 10% 이하인 경우 K-S 통계량은 $N=500$일 때 0.9268이고, $N=50{,}000$일 때 0.9279로 큰 차이가 없다. 반면에 동일한 조건에서 제II종오류율이 60%인 경우 K-S 통계량은 $N=500$일 때 0.4526이고, $N=50{,}000$일 때 0.4130으로 표본크기가 커짐에 따라서 K-S 통계량의 차이가 커지는 것을 알 수 있다. $p=0.05$인 경

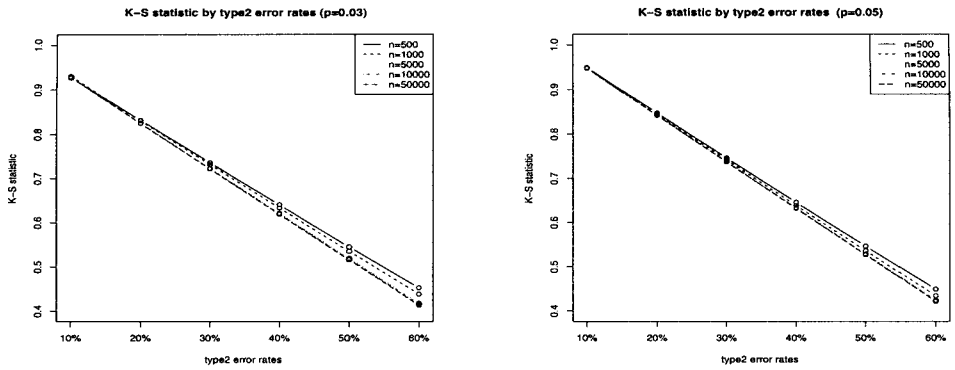

그림 3.1: $p = 0.03$과 $p = 0.05$에 대한 제II종오류율별 K-S 통계량

우에는 $p = 0.03$보다 차이가 작지만 표본의 크기가 커짐에 따라서 차이가 커지는 형태는 $p = 0.03$과 동일하다.

표 3.1을 통해 얻어진 그림 3.1은 $p = 0.03$과 $p = 0.05$에 대한 제II종오류율별 K-S 통계량을 나타낸다. 그림 3.1의 왼쪽그림은 $p = 0.03$의 결과를 나타낸다. $p = 0.03$의 경우 제II종오류율이 증가함에 따라서 K-S 통계량은 선형적으로 감소한다. 제II종오류율이 작은 경우 표본크기에 따른 K-S 통계량의 차이는 거의 없다. 제II종오류율이 증가함에 따라서 표본크기가 큰 경우가 표본크기가 작은 경우보다 K-S 통계량의 감소량이 큰 것을 알 수 있다. 그림 3.1의 오른쪽 그림은 $p = 0.05$에서 표본크기별, 제II종오류율별 K-S 통계량을 나타낸다. K-S 통계량은 역시 선형적으로 감소하고 $p = 0.03$과 유사하며 제II종오류율이 커짐에 따른 표본크기별 변화는 $p = 0.03$인 경우보다 약간 작다고 판단된다.

3.2. 대안적인 판단기준

표본크기와 불량률 그리고 제II종오류율을 고려하여 모의실험된 K-S 통계량의 결과는 3.1절과 같다. 이를 Joseph (2005)가 제안한 표 1.1과 비교하고 대안이 되는 판단 방법에 대해서 살펴보자. 먼저 표본크기 $N \geq 5{,}000$인 대표본인 경우의 K-S 통계량 결과를 기준으로 살펴보면 대표본에서 K-S 통계량은 소표본보다 작은 값을 나타내는데, 제II종오류율이 10% 이하인 경우에는 0.92 이상으로 그리고 제II종오류율이 60%인 경우에는 그보다 작은 값이지만, 일반적으로 적용되고 있는 0.30~0.40 수준보다는 큰 값인 0.41 이상의 큰 값을 나타낸다. 그러므로 Joseph (2005)가 제안한 표 1.1의 판단기준에 의하면 어떤 모형이라도 'Satisfactory'라는 판단을 내린다. 하지만 표 3.1에서 보는 것처럼 K-S 통계량이 0.4보다 크더라도 제II종오류율이 60% 이상 될 수도 있다. 이와 같은 결과는 제II종오류율의 관점에서 보았을 때 잘못된 판단으로 해석될 수 있다. 즉 제II종 오류율이 30 또는 40% 정도 되는 모형을 판별력이 좋은 모형이라고 판단하고자 한다면 K-S 통계량의 기준은 좀더 상향될 필요가 있다. 또 다른 관점에서 신용평가모형의 사용자의 관점에서 제II종오류율이 크더라도 전체 기대비용 면에서 허용할 수 있다면 해당 모형을 채택할 수 있을 것

이다. 이러한 점을 반영하기위해 앞서 설명한 모의실험에서 얻어진 표본크기와 제II종오류율을 기준으로한 K-S 통계량의 상위 90과 95백분위수 u_{90}과 u_{95}를 임계기준으로 새로운 판단기준을 제안하였다.

표 3.1의 결과를 통해 살펴보면 모의실험된 K-S 통계량의 90과 95백분위수의 값은 표본크기에 영향을 받는다. $p=0.03$에 대한 제II종오류율 30%의 95백분위수 결과를 살펴보자. $N=500$인 경우 0.7869이고 $N=50{,}000$이면 0.7251을 나타내는데 표본크기가 커질수록 95백분위수는 작아진다고 파악된다. $p=0.05$에 대한 제II종오류율 40%의 90백분위수 결과는 $N=500$인 경우 0.6786이고 $N=50{,}000$이면 0.6345를 나타낸다. 불량률이 달라지고 표본크기가 달라질수록 90과 95백분위수는 달라진다. 따라서 판단기준 값은 불량률과 표본크기를 함께 고려할 필요가 있다. 표 3.1에 제시된 K-S 통계량의 상위 90과 95백분위수인 u_{90}과 u_{95}를 임계기준으로 새로운 판단방법으로 제안하는 가설검정 방법을 수행해보자.

예를 들어 불량률 $p=0.03$이고 표본크기 $N=5{,}000$에서 제II종오류율 40%에서의 95백분위수는 0.6339이다. 95백분위수를 유의수준 5%에서의 임계값으로 간주하여 평가모형에서 생성된 K-S 통계량 값이 0.60이면 두 집단이 동일하다는 귀무가설 $H_0 : F_1(x) = F_2(x)$을 유의수준 5%에서 기각할 수 없다. 반면에 제II종오류율을 50%까지 허용할 수 있다면 모형은 제II종오류율 50% 수준에서 귀무가설을 기각하여 모형이 적합하다고 판단할 수 있다.

또 하나의 예를 들어보자. 구축된 신용평가모형에서 $p=0.03$이고 표본크기 $N=1{,}000$이며 산출된 K-S 통계량이 0.70이라고 하자. 이러한 경우 K-S 통계량은 단순하게 매우 큰 값으로 판별력이 좋은 모형으로 판단할 수 있다. 하지만 제II종오류율을 20%까지밖에 허용할 수 없다면 이는 역시 귀무가설을 기각하지 못하게 되어 K-S 통계량 값은 0.70이라 하더라도 적합한 모형으로 채택될 수 없다.

홍종선 등 (2008)에 수록되어 있는 실제 사례를 대상으로 표 3.1을 이용한 신용평가모형의 판별력에 대한 검증을 해보자. 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 자료이다. 표본크기는 4,268건으로 표 3.1에 대응하여 근사값인 $N=5{,}000$을 이용한다. 불량률은 모형생성을 위해 수집된 전체자료 중에서 불량이 차지하는 비율로써 $\hat{p}=167/4{,}268=0.039$이므로 $p=0.03$의 기준을 적용하자. 실제 신용평가모형에서 산출된 K-S 통계량은 0.667이므로 표 3.1의 $N=5{,}000$에서 유의수준 5%에서 제II종오류율 30% 정도의 모형을 고려한다면 임계기준은 0.7344로 산출된 모형은 두 집단이 동일하다는 귀무가설을 기각할수 없게 된다. 반면에 제II종오류율 40%정도 수준의 모형을 고려한다면 임계기준이 0.6339로 유의수준 5%에서 기각하여 판별력이 있는 모형이라고 결론내릴 수 있다.

4. 결론

일반적인 판단기준인 Joseph (2005)의 표 1.1을 일괄적으로 적용하는 데서 발생하는 문제점을 개선하기 위하여, 본 연구에서는 현재 신용평가(credit rating) 현장에서 이루어지고 있는 환경과 유사하게 설정한 모의실험을 통하여 생성된 K-S 통계량과 표 1.1의 기준을 비교하여보았다. 표본크기(N), 불량률(p) 그리고 제II종오류율의 변화에 따른 자료를 생성해서 두 분포함수의 동일성을 검정하는 K-S 통계량의 변화를 살펴보았다. 불량률은 $p=0.03$과 $p=0.05$를 고려하였고 표본크기는 500, 1,000, 5,000, 10,000, 50,000을 살펴보았는데, 제II종오류율이 커짐에 따라 K-S 통계량은 감소하고 표본크기가 커질수록 더 작은 값을 갖는다. 불량률 $p=0.05$보다 $p=0.03$이 표본크기가 커짐에 따라서 작은 값을 나타내고 작은 제II종오류율에서는 불량률에 따라 K-S 통계량의 차이가 크게 나타나지만 제II종오류율이 커짐에 따라서 K-S 통계량의 차이는 크지 않았다. Joseph (2005)가 제안한 표 1.1의 판단기준에 대한 대안으로 표본크기 N과 불량률 p 그리고 제II종오류율 정보에 근거한 대안적인 판단기준 상위 90과 95백분위수를 제시하였다. 대안적인 판단기준은 상위 90과 95백분위수를 기준으로 허용할수 있는 제II종오류 범위 내에서 유의수준 10%와 5%에서 가설검정할 수 있다.

표본크기(N), 불량률(p)은 미리알고 있지만 분류기준 스코어 x_c에 따라 변하는 오류율은 사전에 알 수 없다. 또한 신용평가모형 사용자에 따라 허용할 수 있는 제II종오류율의 범위는 다를 수 있다. 실제 사례에서 제I종오류율과 제II종오류율은 알 수 없지만 표 3.1의 K-S 통계량 결과와 신용평가모형의 K-S 통계량을 이용하면 제I종오류율이 5%일 때 두 분포함수의 동일성을 검정하는 귀무가설을 기각할 수 있는 제II종오류율의 허용범위를 파악할 수 있다. 그러므로 신용평가모형 사용자가 2절에서 언급한 기대비용함수를 고려하여 제II종오류율을 설정하면 신용평가모형의 가설검정이 가능하다.

본 연구에서는 제I종오류율을 5%로 고정하고 다양한 제II종오류율을 갖는 자료를 생성하여 K-S 통계량을 산출하고 대안적인 판단기준으로 표 3.1을 제시하였다. 표 3.1은 표본크기, 불량률, 제II종오류율을 고려한 K-S 통계량의 판단기준으로 적용할 수 있으며 표 1.1 보다 개선된 판단기준이라고 할 수 있다.

감사의 글

본 연구과정에서 조언을 아끼지 않은 한국기업평가의 임한승 박사와 한국개인신용의 홍성식 박사에게 감사합니다.

참고문헌

송문섭, 박창순, 이정진 (2003). <S-LINK를 이용한 비모수통계학>, 자유아카데미.
홍종선, 이창혁, 김지훈 (2008). 범주형 재무자료에 대한 신용평가모형 검증 비교, <한국통계학회논문집>, **15**, 615–631.
Daniel, W. W. (1990). *Applied Nonparametric Statistics,* 2nd ed, PWS-Kent, Boston.

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

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

Fernandes, J. E. (2005). Corporate credit risk modeling: Quantitative rating system and probability of default estimation. Available from : http://pwp.netcabo.pt/jed fernandes/JEF CorporateCreditRisk.pdf.

Hand, D. J. (1994). Assessing classification rules, *Journal of applied statistics*, **21**, 3-16.

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

Koh, H. C. (1992). The sensitivity of optimal cutoff points to misclassification costs of Type I and Type II errors in the going concern prediction context, *Journal of Business Finance & Accounting*, **19**, 187-197.

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

Thomas, L. C., Edelman, D. B. and Crook, J. B. (2002). *Credit Scoring and Its Applications*, Society for Industrial Mathematics, Philadelphia.

Wilkie, A. D. (1992). Measures for comparing scoring systems, *Credit Scoring and Credit Control*, Eds. Thomas, L. C., Crook, J. N and Edelman, D. B. Oxford: Carendon, 123-138.

Wilkie, A. D. (2004). Measures for comparing scoring systems, In *Readings in Credit Scoring-recent developments, advances, and aims*, Eds. Thomas, L. C., Crook, J. N, and Edelman, D. B. Oxford finance.

[2008년 9월 접수, 2008년 10월 채택]

Some Issues on Criterion for Kolmogorov-Smirnov Test in Credit Rating Model Validation

Yong Seok Park[1], Chong Sun Hong[2]

Abstract

Kolmogorov-Smirnov(K-S) statistic has been widely used for the model validation of credit rating models. Validation criteria for the K-S statistic is empirically used at the levels of 0.3 or 0.4 which are much larger than the critical values of K-S test statistic. We examine whether these criteria are reasonable and appropriate through the simulations according to various sample sizes, type II error rates, and the ratio of bads among data. The simulation results say that the currently used validation criteria are too lower than values of K-S statistics obtained from any credit rating models in Korea, so that any credit rating models have good discriminatory power. In this work, alternative criteria of K-S statistic are proposed as critical levels under realistic situations of credit rating models.

Keywords: Credit rating model; critical value; discriminatory power; Kolmogorov-Smirnov statistic; validation.

1) Researcher, Research Institute of Applied Statistics, Sungkyunkwan University, Seoul 110-745, Korea.

2) Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea. Correspondence: cshong@skku.ac.kr