응용통계연구 (2008)
21(6), 1065–1075

신용평가를 위한 Kolmogorov-Smirnov 수정통계량

홍종선[1] · 방글[2]

[1]성균관대학교 통계학전공; [2]성균관대학교 응용통계연구소

(2008년 8월 접수, 2008년 10월 채택)

요 약

신용평가모형 개발과 적합성 검정 연구에서 부도율분포로부터 부도기업과 정상기업의 판별력을 검정하는 방법으로 비모수적인 방법인 Kolmogorov-Smirnov(K-S) 검정방법을 많이 사용한다. 모집단에 대한 누적분포함수를 알고있으며 이 분포함수가 두 개의 분포함수로 분할되었다는 가정하에서 두 분포함수의 동일성을 검정하는 신용평가 연구에서 스코어 또는 부도율이 다양한 확률분포를 따른다고 가정하고 기존의 K-S 통계량과 수정된 K-S 통계량을 비교 토론한다.

주요용어: 리스크, 부도율, 분포함수, 비모수검정, 신용평가, 스코어, 판별력.

1. 서론

확률표본 $\{X_1, X_2, \ldots, X_{n+m}\}$의 확률밀도함수 $f(x)$가 다음과 같이 확률밀도함수 $f^1(x)$와 $f^2(x)$로 분할되었다고 가정하자.

$$f(x) = \alpha f^1(x) + (1-\alpha) f^2(x), \tag{1.1}$$

여기서 $\alpha \in (0, 1)$이며 $x \in (-\infty, \infty)$이다. 이와 같은 상황은 금융시장에서 리스크 관리를 위해 신용평가모형을 개발하고 적합성 검정(validation)할 때에 부도율(Probability of Default: PD)에 관한 판별력(discriminatory power)을 탐색하는 연구에서 많이 찾아볼 수 있다. 확률변수 X를 스코어(score)라 하고, $f^1(x)$와 $f^2(x)$를 각각 $f_D(x)$와 $f_N(x)$로 정의하여 부도(default)기업과 정상(non-default)기업의 스코어 확률밀도함수를 나타낸다고 가정하자. 스코어는 부도율을 변수변환하여 점수형태로 표현하기 때문에 스코어분포(distribution of score)와 부도율분포(distribution of PD)는 유사하게 사용된다. 또한 스코어의 누적분포함수를 $F(x)$로 정의하고 부도기업과 정상기업의 누적분포함수를 $F_D(x)$와 $F_N(x)$로 정의하면, 스코어의 누적분포함수는 다음과 같이 분할된다.

$$F(x) = \alpha F_D(x) + (1-\alpha) F_N(x), \tag{1.2}$$

여기에서 α는 부도율 총합(total (portfolio-wide) PD)이다.

Tasche (2006)는 스코어 또는 부도율과 유사한 자료에서 판별력을 탐색하는데 유용한 그래픽적 방법인 ROC와 CAP함수를 식 (1.2)을 이용하여 다음과 같이 정의하였다. $u \in (0, 1)$에 대하여,

$$\mathrm{ROC}(u) = F_D\left(F_N^{-1}(u)\right), \qquad \mathrm{CAP}(u) = F_D\left(F^{-1}(u)\right).$$

[1]교신저자: (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 경제학부 통계학전공, 교수.
E-mail: cshong@skku.ac.kr
[2](110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 응용통계연구소, 연구원.
E-mail: bgwhite@skku.edu

또는 스코어 $s \in R$에 대응하는 모든점 $(F_N(s), F_D(s))$와 $(F(s), F_D(s))$을 연결함으로 ROC곡선과 CAP곡선을 각각 표현한다.

본 연구에서는 스코어의 누적분포함수가 부도와 정상기업의 누적분포함수로 분할되었다는 가정 하에 다음과 같은 가설을 고려하자.

$$H_0 : F_D(\cdot) = F_N(\cdot)$$
$$H_1 : F_D(\cdot) > F_N(\cdot).$$

신용평가 연구에서 식 (1.2)와 같이 분할된 부도기업과 정상기업의 스코어분포에 관한 동일성 검정방법으로는 비모수적인 방법인 이표본 Kolmogorov-Smirnov(K-S) 검정법을 사용한다 (Smirnov, 1939; Darling, 1957; Barton과 Mallows, 1965; Hájek 등, 1998 참조). K-S 검정방법을 본 연구의 경우로 정리하면 다음과 같다. $\{X_1, X_2, X_3, \ldots, X_{n+m}\}$을 식 (1.1)과 (1.2)에서 언급한 $f(\cdot)$, $F(\cdot)$로부터 추출한 크기 $n+m$의 확률표본이라고 하자. 그 중에서 크기 n개의 확률표본은 부도기업의 부도율분포인 $F_D(\cdot)$로부터 추출하고, 크기 m개의 확률표본은 정상기업의 부도율분포인 $F_N(\cdot)$로부터 추출한다. 확률표본에 대한 분포와 확률을 정리한 표 2.1을 참조하면, K-S 검정통계량은 다음과 같이 정의된다.

$$\text{K-S} = \max_x \left| \hat{F}_D(x) - \hat{F}_N(x) \right|, \tag{1.3}$$

여기서 각 확률표본에 대한 표본분포함수(sample distribution function)는 다음과 같다:

$$\hat{F}_D(x) = \frac{1}{n} \sum_{i=1}^{n+m} I(X_i \le x | D = 1), \qquad \hat{F}_N(x) = \frac{1}{m} \sum_{i=1}^{n+m} I(X_i \le x | D = 0).$$

K-S 검정방법은 K-S 통계량의 분포표를 이용하여 귀무가설을 기각한다. 소표본인 경우에는 각 표본수에 따라 유의수준별 임계값(critical value)을 사용하고, 대표본인 경우에는 표본크기의 함수로 나타나는 임계값을 사용한다. 예를 들어 유의수준이 5%이고 대표본인 경우에는 $1.22\sqrt{(n+m)/nm}$인 임계값을 사용한다 (Daniel, 1990; 송문섭 등, 2003 참조). 표본수를 4,000이라고 할 때의 유의수준 5%에서 K-S검정통계량의 임계값은 0.08정도로 작은 값을 갖는다. 그러나 신용평가 연구에서 대부분의 K-S 통계량은 매우 큰 값으로 나타난다. 따라서 일반적인 모형의 검정기준을 사용하는데 Joseph (2005)가 제안한 K-S 통계량 검정기준은 K-S 통계량이 0.38이상이면 'Satisfactory', 0.47이상이면 'Good' 그리고 0.55이상이면 'Very Good'이라고 판정한다. 대표본인 경우에 통계학적인 K-S 통계량의 임계값은 매우 작은 값을 가지는 반면에 신용평가 연구에서의 K-S 통계량은 대부분 0.6이상의 매우 큰 값을 나타내어 통계학적으로는 항상 유의한(significant) 결과이므로 표본의 크기와 무관한 일반적인 법칙(Rule of Thumb)을 사용한다.

K-S 검정방법은 두 모집단 분포함수의 동일성을 검정한다. 그러나 신용평가 연구에서는 식 (1.1)과 (1.2)에서와 같이 하나의 모집단의 확률밀도함수와 누적분포함수 $f(\cdot)$, $F(\cdot)$를 두 개의 분포함수로 분할하여 두 분포함수의 동일성을 검정한다는 차이가 있다. 그리고 신용평가 연구에서 사용하는 대부분의 자료는 양의 왜도를 갖는 분포형태를 따른다는 것을 알고있다. 본 연구에서는 모집단의 분포형태를 경험적으로 파악할 수 있거나 사전에 알고 있는 경우를 가정한다. 식 (1.1)과 (1.2)에서 정의한 확률밀도함수나 누적분포함수를 사전에 알고 있는 경우 또는 최소한 추정이 가능하다는 가정 하에 분할된 두 누적분포함수 $F_D(\cdot)$와 $F_N(\cdot)$의 동일성에 대한 검정방법을 연구하고자 한다 .

K-S 검정통계량은 분포와 무관(distribution free)하다는 전제 하에서 표본분포함수를 사용하지만, 본 연구에서는 확률밀도함수를 알고 있다는 가정 하에 누적분포함수를 이용하여 K-S 통계량을 수정하고자 한다. 기존의 K-S 통계량에 사용하는 분할된 표본분포함수 대신 누적분포함수의 추정량으로 정의하여

표 2.1. K-S 수정통계량 계산

ID	X	D	$P_D(x)$	$P_N(x)$	$\hat{F}_D^M(x)$	$\hat{F}_N^M(x)$
1	x_1	1				
2	x_2	1				
3	x_3	0				
4	x_4	1	$\vdots$	$\vdots$	$\vdots$	$\vdots$
5	x_5	1	$P_D(x_i) \equiv$	$P_N(x_i) \equiv$	$\hat{F}_D^M(x_i) \equiv$	$\hat{F}_N^M(x_i) \equiv$
$\vdots$	$\vdots$	$\vdots$	$P(X = x_i \mid D = 1)$	$P(X = x_i \mid D = 0)$	$P(X \le x_i \mid D = 1)$	$P(X \le x_i \mid D = 0)$
		0				
$\vdots$	$\vdots$	1	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	x_N	0			1	1
$N = m + n$		n	1	1		

K-S 수정통계량(modified K-S statistic)을 2절에서 제안하고, 간단한 예제를 통해 K-S 수정통계량을 이해한다. 3절에서는 신용평가에 관한 연구에서 부도율을 따르는 다양한 확률분포인 경우를 예를 들어 살펴본다. 정규분포를 고려한 후에, 대부분의 부도율분포는 오른쪽으로 꼬리가 길게 늘어져 있기 때문에 치우친 정규(skew normal)분포와 베타(Beta)분포에서 양의 왜도를 갖는 경우의 K-S 수정통계량을 구하고 K-S 통계량과 비교 분석한다. K-S 수정통계량과의 비교 및 결론은 4절에서 논의한다.

2. K-S 수정통계량

2.1. K-S 수정통계량

스코어의 확률밀도함수(probability density function) 또는 확률질량함수(probability mass function)를 $P(X = x)$라고 하면, $P_D(X = x)$와 $P_N(X = x)$는 부도와 정상일 때 각각의 확률질량함수이며 다음과 같이 조건부 확률질량함수(conditional probability mass function)로 정의한다.

$$\begin{aligned} P_D(X = x) &\equiv P(X = x \mid D = 1), \\ P_N(X = x) &\equiv P(X = x \mid D = 0), \end{aligned} \tag{2.1}$$

여기서 확률변수 D는 부도상태를 나타내는 지시함수(indicator function)이며, 부도율 총합 α는 $P(D = 1)$이다. 따라서 식 (1.1)은 다음과 같이 표현된다.

$$\begin{aligned} P(X = x) &= \alpha P_D(X = x) + (1 - \alpha) P_N(X = x) \\ &= \alpha P(X = x \mid D = 1) + (1 - \alpha) P(X = x \mid D = 0). \end{aligned}$$

그리고 $F_D^M(x)$와 $F_N^M(x)$는 부도와 정상일 때 각각의 누적분포함수(cumulative distribution function)이며 다음과 같이 정의한다.

$$\begin{aligned} F_D^M(x) &\equiv P(X \le x \mid D = 1), \\ F_N^M(x) &\equiv P(X \le x \mid D = 0). \end{aligned} \tag{2.2}$$

식 (1.2)는 다음과 같이 누적분포함수의 분할로 표현된다.

$$F(x) = \alpha F_D^M(x) + (1 - \alpha) F_N^M(x).$$

표 **2.2.** K-S 통계량

ID	X	D	$P(x)$	$P_D(x)$	$P_N(x)$	$\hat{F}_D(x)$	$\hat{F}_N(x)$	차이
1	−1.8	1	0.1	0.333	0.000	0.033	0.000	0.0330
2	−1.5	0	0.1	0.000	0.143	0.033	0.143	0.1100
3	−1.2	1	0.1	0.333	0.000	0.667	0.143	0.5240
4	−0.8	0	0.1	0.000	0.143	0.667	0.286	0.3810
5	−0.5	1	0.1	0.333	0.000	1.000	0.286	0.7140⋆
6	0.0	0	0.1	0.000	0.143	1.000	0.429	0.5710
7	0.5	0	0.1	0.000	0.143	1.000	0.571	0.4290
8	0.8	0	0.1	0.000	0.143	1.000	0.714	0.2860
9	1.2	0	0.1	0.000	0.143	1.000	0.857	0.1430
10	1.5	0	0.1	0.000	0.143	1.000	1.000	0.0000
		3		1	1			

표 **2.3.** K-S 수정통계량

ID	X	D	$P(x)$	$P_D(x)$	$P_N(x)$	$\hat{F}_D^M(x)$	$\hat{F}_N^M(x)$	차이
1	−1.8	1	0.0359	0.1986	0.0000	0.1986	0.0000	0.1986
2	−1.5	0	0.0309	0.0000	0.0377	0.1986	0.0377	0.1610
3	−1.2	1	0.0483	0.2668	0.0000	0.4655	0.0377	0.4278
4	−0.8	0	0.0968	0.0000	0.1182	0.4655	0.1559	0.3096
5	−0.5	1	0.0967	0.5345	0.0000	1.0000	0.1559	0.8441⋆
6	0.0	0	0.1915	0.0000	0.2337	1.0000	0.3896	0.6104
7	0.5	0	0.1915	0.0000	0.2337	1.0000	0.6233	0.3767
8	0.8	0	0.0967	0.0000	0.1180	1.0000	0.7414	0.2586
9	1.2	0	0.0968	0.0000	0.1182	1.0000	0.8595	0.1405
10	1.5	0	0.1151	0.0000	0.1405	1.0000	1.0000	0.0000
		3		1	1			

식 (1.3)에서 언급하였듯이 $F_D(\cdot)$와 $F_N(\cdot)$의 표본분포함수 $\hat{F}_D(x)$와 $\hat{F}_N(x)$를 이용하여 비모수적인 K-S 검정방법을 사용하는 대신에 알려진 분포함수의 분할된 분포함수의 추정분포함수를 이용하여 K-S 수정통계량(modified K-S statistic)을 다음과 같이 제안한다.

$$\text{modified K-S} = \max_x \left| \hat{F}_D^M(x) - \hat{F}_N^M(x) \right|, \tag{2.3}$$

여기서 $\hat{F}_D^M(x)$와 $\hat{F}_N^M(x)$은 식 (2.2)의 추정분포함수이다. 예를 들어 스코어 자료와 이에 대응하는 부도상황을 나타내는 자료를 다음과 같은 표 2.1의 형태로 작성하면, 부도와 정상일 때 각각의 확률질량함수 $P_D(X = x)$와 $P_N(X = x)$를 구할 수 있으며 부도와 정상일 때 각각의 누적분포함수 $F_D^M(x)$와 $F_N^M(x)$를 추정하는 $\hat{F}_D^M(x)$와 $\hat{F}_N^M(x)$를 얻을 수 있다.

2.2. 예제

식 (1.1)의 확률밀도함수의 $f(x)$와 누적분포함수 $F(x)$를 표준정규분포의 $\phi(x)$와 $\Phi(x)$로 가정하자. 추출한 10개의 확률표본을 스코어로 간주하고, 그 중에서 부도기업을 3개로 설정하였다. K-S 통계량을 구하기 위한 표본분포함수 $\hat{F}_D(x)$와 $\hat{F}_N(x)$를 표 2.2에, K-S 수정통계량을 구하기 위한 추정분포함수 $\hat{F}_D^M(x)$와 $\hat{F}_N^M(x)$를 표 2.3에 각각 나타내었다. 표 2.2에서 식 (2.1)의 조건부 확률질량함수 $P_D(x)$와

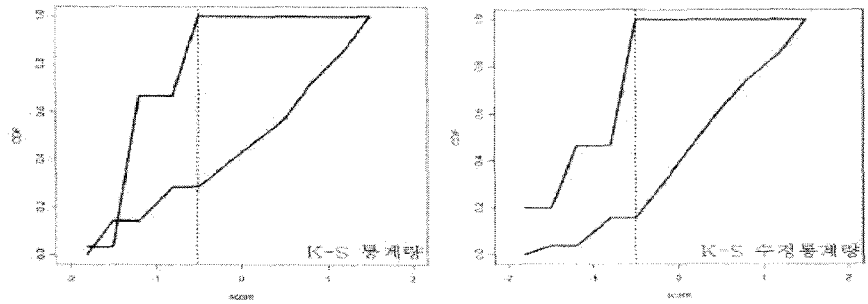

그림 2.1. K-S 통계량과 K-S 수정통계량

$P_N(x)$는 균일분포를 따르므로 일정하지만 표 2.3의 조건부 확률질량함수는 정규분포를 따르기 때문에 크기가 일정하지 않다.

K-S 통계량과 K-S 수정통계량이 결정되는 스코어는 5번째인 -0.5에서 발생하지만 각각 그 값은 다르며 K-S 수정통계량이 0.8441로 0.13정도 크다. 부도기업과 정상기업의 표본분포함수와 추정분포함수를 그림 2.1에 각각 나타내었다. 두 분포함수의 최대차이를 참조선(reference line)으로 같이 표현하였다.

3. 다양한 확률분포의 예제

3.1. 정규분포

신용평가모형 개발을 위해서 사용하는 자료와 유사한 상황을 설정하고 작은 표본을 생성하여 K-S 통계량과 K-S 수정통계량을 비교하고자 한다. 우선 표준정규분포로부터 50개의 난수를 추출하여 스코어라고 간주하자. 일반적으로 스코어는 부도확률의 순위에 근거하여 변수변환한 점수형태로 나타나지만 여기서는 실수의 값을 부도율로 가정한다. 모집단의 분포형태를 사전에 알고 있으며 전체 분포함수가 부도기업과 정상기업의 분포함수로 분할되었다고 가정하고, 두 분포함수의 동일성을 검정하는 K-S 수정통계량을 제안해 보았다. K-S 수정통계량에 표준정규분포를 적용하고, 실제 부도율 분포와 유사한 치우친 정규분포와 베타분포에서 양의 왜도를 갖는 경우를 50개 중에서 90%를 실제 정상기업이라 하고, 나머지 10%를 실제 부도기업이라고 가정한다(실제로는 정상기업의 비율이 95% 이상이나 본 연구에서는 90%로 설정한다). 그리고 정상기업을 부도기업으로 예측하는 오분류율을 5%, 10%, 15%, 20%의 네 가지 경우로 하고, 부도기업을 정상기업으로 예측하는 비율은 적은 비율인 5%로 고정한다(일반적으로 부도기업을 정상으로 예측하는 것보다 정상기업을 부도기업으로 예측하는데 관심이 많기 때문임). 실제 신용평가모형 개발에 사용하는 자료를 보면 실제자료와 예측자료의 차이가 발생하는 경우 즉 정상기업을 부도기업으로 부도기업을 정상기업으로 잘못 예측되는 경우는 정상과 부도라고 구분하는 판별기준에 가까운 곳에서 자주 발생한다. 판별기준과 멀리 떨어진 경우에는 잘못 예측되는 경우가 발생하지 않는다. 이러한 실제자료의 특성과 유사한 형태를 만들어주기 위하여 정상기업을 부도기업으로 예측하는 오분류율을 가중치를 주어 적용하였다. 우선 정상기업이라고 간주한 90%의 자료를 5개의 등구간으로 나누어 판별기준에서 가까운 구간부터 오분류율 비율을 6:3:1:0:0으로 적용하여, 전체 오분류된 자료가 멀리 떨어진 두 구간에서는 발생하지 않으며 가까운 곳에서는 많이 발생하도록 표본을 생성한다. 이렇게 생성된 자료로 식 (2.3)을 적용한 K-S 수정통계량을 구한다. 이 과정을 5,000번 반복하여 구한 K-S 통계량과 정규분포의 가정 하에서 K-S 수정통계량의 평균값을 표 3.1에 나타내었다.

표 3.1. 표준정규분포에서 K-S 수정통계량

오분류율	K-S	수정 K-S	차이
5%	0.7523	0.7958	−0.0435
10%	0.6632	0.7119	−0.0488
15%	0.6567	0.7064	−0.0498
20%	0.6778	0.7195	−0.0417

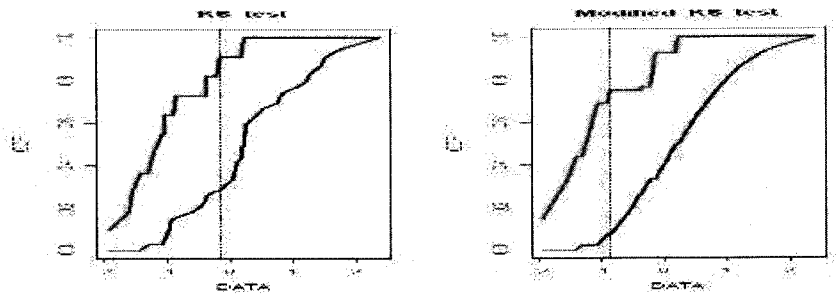

그림 3.1. 표준정규분포에서 K-S 통계량(오분류율 15%)

모집단이 정규분포를 따르는 경우에 기존의 K-S 통계량과 비교하여 K-S 수정통계량은 조금 큰 값을 가지며 오분류율 15%일 때 최대 5%정도의 차이를 갖는다. 오분류율이 커질수록 K-S 통계량과 K-S 수정통계량은 모두 작아지는데, 20%에서 값이 조금 커지는 역전 현상이 발생한다. 이것은 오분류율의 증가로 정상기업 분포 중 판별기준 가까운 곳에서 정상기업보다 부도기업의 수가 커져서, 오분류율이 큰 경우에 정상기업과 부도기업을 90 : 10으로 가정한 비율에서 벗어나기 때문이다.

Joseph (2005)가 제안한 신용평가모형의 적합성 검정기준 중 K-S 통계량이 0.6184이상이면 'Strong', 0.6827이상이면 'Very Strong' 그리고 0.7394이상이면 'Excellent'라고 판정하는데, 표 3.1에서 K-S 통계량과 K-S 수정통계량의 값을 바탕으로 모형은 'Strong'이상이라고 판단내릴 수 있다. 특히 K-S 통계량에서는 'Very Strong'이 오분류율 5%에서만 발생하지만 K-S 수정통계량에서는 5%∼20%의 모든 오분류율에서 'Very Strong'한 결과를 보이는 것을 확인할 수 있다.

모의실험한 정규분포의 자료(오분류율 15%)를 바탕으로 식 (1.3)의 K-S 통계량에 사용하는 표본분포함수와 식 (2.3)의 K-S 수정통계량을 정의하는데 사용하는 식 (2.2)의 추정분포함수를 각각 그림 3.1에 표현하였다. 각각의 그림에서 위의 함수가 부도기업의 분포함수이며, 아래 함수는 정상기업의 분포함수이다. 각 그림에서 참조선은 두 분포 함수의 차이가 최대일 때를 나타낸다. 두 그림을 비교하면, 표본분포함수와 추정분포함수의 차이가 많이 나는 것을 파악할 수 있다. 분포함수의 곡선 형태는 두 그림 모두 부도기업 보다는 정상기업의 수가 많기 때문에 정상기업의 곡선이 완만하게 증가하는 형태이다. 부도기업의 표본분포함수보다 추정분포함수의 증가폭이 일정하지 않고 스코어가 증가할수록 커지며, 정상기업의 표본분포함수보다 추정분포함수가 부드럽고 단조롭게 증가한다.

3.2. 양의 왜도를 갖는 치우친 정규분포

치우친 정규분포(skew normal distribution)는 수학적 전개와 표준정규분포의 적용에 대하여 활발히 연구된 문헌들을 많이 발견할 수 있다. 치우친 정규분포의 특성들에 대하여 수학적으로 논의한 연구로는 Azzalini (1985), Chiogna (1998), Henze (1986)가 있으며, 특히 통계적으로 치우친 분포 형태에 대

표 **3.2.** 치우친 정규분포에서 K-S 수정통계량

λ	오분류율	K-S	수정 K-S	차이
2 왜도: 0.4538	5%	0.7553	0.7951	−0.0398
	10%	0.6644	0.7067	−0.0423
	15%	0.6560	0.6990	−0.0430
	20%	0.6797	0.7140	−0.0344
5 왜도: 0.8510	5%	0.7543	0.7944	−0.0402
	10%	0.6651	0.7075	−0.0424
	15%	0.6567	0.6990	−0.0423
	20%	0.6797	0.7140	−0.0344

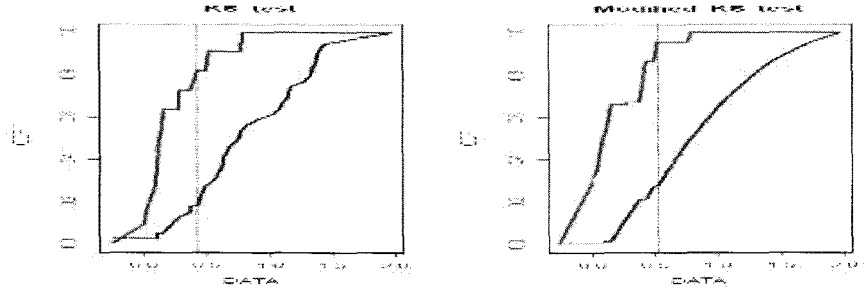

그림 3.2. 치우친 정규분포 $SN(5)$에서 K-S 통계량(오분류율 15%)

한 연구로는 Azzalini와 Capitanio (1999), Salvan (1986), Liseo (1990), Chang 등 (2002), Genton (2005)이 있다. Gupta와 Chen (2001)은 치우친 정규분포의 누적분포함수 표를 제안하고 치우친 정규분포의 적합성검정에 이용하였다.

대부분의 부도율분포(distribution of PD) 또는 스코어분포는 큰 양수 값의 왜도를 갖는다. 그러므로 치우친 정규분포(skew normal)에서 양의 왜도를 갖는 경우를 고려하자 (Azzalini, 1985; Azzalini와 Capitanio, 1999 참조). 치우친 정규분포($SN(\lambda)$)로부터 50개의 난수를 추출하고 90 : 10의 비율로 실제 정상기업과 부도기업의 스코어를 생성한다. 여기서 λ는 왜도계수로 정의한다 (Buccianti, 2005; Gupta 등, 2004 참조). 정상기업을 부도기업으로 예측하는 오분류율은 5%, 10%, 15%, 20%의 네 가지 경우로 하고, 부도기업을 정상기업으로 예측하는 비율은 5%로 3.1절과 같이 설정한다. 왜도계수 λ는 2와 5인 경우를 고려하였고 각각 대응하는 왜도는 0.4538과 0.8510이다. λ의 스코어에 따른 누적확률값은 Gupta와 Chen (2001)의 치우친 정규분포 표에 의해 구한다. 이 과정을 5,000번 반복하여 구한 K-S 통계량과 치우친 정규분포에서 양의 왜도를 갖는 경우에서 K-S 수정통계량의 평균값을 표 3.2에 나타내었다.

치우친 정규분포에서 양의 왜도를 갖는 경우에 K-S 통계량은 λ값에 관계없이 오분류율에 의존하며 그 값은 정규분포 가정 하에서의 표 3.1과 매우 유사함을 발견할 수 있다. K-S 수정통계량은 λ가 커짐에 따라 즉 왜도가 커짐에 따라 조금 작은 값을 가지며, 오분류율이 증가할수록 그 차이는 없어진다. 또한 K-S 수정통계량은 정규분포 가정 하에서와 비교하여 조금 작은 값을 가지므로 K-S 통계량과 비교하였을때 정규분포의 경우보다 조금 작은 차이를 보인다.

모의실험한 치우친 정규분포의 자료($\lambda = 5$, 오분류율 15%)를 바탕으로 K-S 통계량의 표본분포함수와 K-S 수정통계량의 추정분포함수를 각각 그림 3.2에 표현하였다. 그림 3.2의 두 그림을 비교하면, 그림

표 3.3. 베타분포에서 K-S 수정통계량

a, b	오분류율	K-S	수정 K-S	차이
$a = 1$	5%	0.7528	0.7968	−0.0439
$b = 5$	10%	0.6621	0.7097	−0.0476
왜도:	15%	0.6551	0.7054	−0.0503
1.2561	20%	0.6786	0.7153	−0.0367
$a = 0.2$	5%	0.7540	0.7979	−0.0438
$b = 5$	10%	0.6648	0.7075	−0.0427
왜도:	15%	0.6557	0.7026	−0.0469
3.4119	20%	0.6796	0.7191	−0.0395

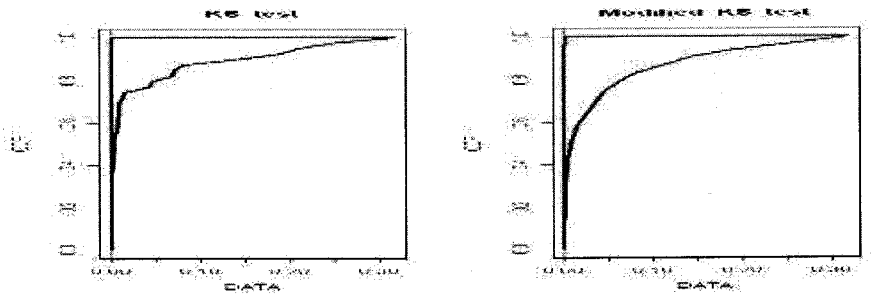

그림 3.3. 베타분포 Beta(0.2, 5)에서 K-S 통계량(오분류율 15%)

3.1과 유사하게 각각의 표본분포함수와 추정분포함수의 차이가 많이 난다는 것을 파악할 수 있다.

3.3. 베타분포

베타분포 Beta(a, b)를 따르는 확률변수는 0과 1사이의 값으로 부도율분포의 확률변수인 부도율(PD)의 성격과 일치한다. 3.2절의 치우친 정규분포와 같이 양의 왜도를 갖는 베타분포 Beta(0.2, 5), Beta(1, 5)로부터 난수를 생성하여 스코어로 간주한다(왜도는 각각 2.6857와 1.0585이다).

각각의 베타분포로부터 50개의 난수를 추출하여, 3.1절과 3.2절에서와 같이 90:10의 비율로 실제 정상기업과 부도기업의 스코어를 생성한다. 그리고 정상기업을 부도기업으로 예측하는 오분류율도 5%, 10%, 15%, 20% 경우로 하고, 부도기업을 정상기업으로 예측하는 비율은 5%로 고정한다. 이 과정을 5,000번 반복하여 구한 K-S 통계량과 베타분포의 가정 하에서 K-S 수정통계량의 평균값을 표 3.3에 나타내었다.

표 3.3에서 K-S 통계량은 베타분포의 모수 a, b에 관계없이 오분류율에 의존하며 그 값은 정규분포 가정 하에서의 표 3.1과 치우친 정규분포에서 양의 왜도를 갖는 경우에서의 표 3.2와 매우 유사함을 발견할 수 있다. K-S 수정통계량은 모수의 값에 따라 왜도가 변해도 큰 차이를 발견할 수 없다. 이것은 정규분포인 경우와 치우친 정규분포와 베타분포에서 양의 왜도를 갖는 경우를 비교했을 때에 약간의 차이가 나는 정도라고 해석된다. 따라서 잘 알려진 K-S 통계량은 분포의 종류와 왜도에 상관없이 오분류율에 따라 다른 값을 갖지만, 본 논문에서 제안하는 K-S 수정통계량은 분포에 따라 다른 값을 갖는다는 것을 발견할 수 있다. 특히 정규분포보다도 양의 왜도를 갖는 분포에서는 오분류율 5%인 경우를 제외하고 조금 작은 값으로 나타난다. 그러나 왜도에 따라 민감하게 반응하지 않는다는 것을 파악할 수 있다.

베타분포 Beta(0.2, 5)의 자료(오분류율 15%)를 바탕으로 K-S 통계량의 표본분포함수와 K-S 수정통계

량의 추정분포함수를 각각 그림 3.3에 표현하였다. 그림 3.1과 3.2와 비교하여 왜도가 심한 이유로 발생 가능 부도기업의 표본분포함수와 추정분포함수는 차이가 없으며, 정상기업의 표본분포함수와 추정분포함수에서 차이를 발견할 수 있다.

4. 결론

적합성 검정(validation)은 신용평가모형의 판별력에 대한 검정뿐만 아니라 금융회사가 스스로 신용평가모형을 관리하고 개선해 나갈 수 있는 능력에 대한 입증과정이라고 볼 수 있어 매우 중요한 문제이다. 특히 비모수적 K-S 검정방법은 적합성 검정방법으로 많이 사용되고 있다. 본 연구에서는 실제 신용평가모형에서 연구하는 모집단의 분포함수를 이미 알고 있으며 전체 분포함수가 부도기업과 정상기업의 분포함수로 분할되었다고 가정하고, 두 분포함수의 동일성을 검정하는 K-S 수정통계량을 제안해 보았다. K-S 수정통계량에 표준정규분포를 적용하고, 실제 부도율 분포와 유사한 치우친 정규분포와 베타분포에서 양의 왜도를 갖는 경우를 적용하여 기존의 K-S 통계량과 비교하였다.

다양한 확률분포 예제의 결과를 살펴보면 스코어가 어떠한 확률분포로부터 생성되어도 K-S 통계량은 오분류율의 변화에 의존한다. 확률분포의 특성과 관계없이 오분류율에 의한 정상기업과 부도기업의 비율만이 반영된 결과임을 알 수 있다. 오분류율의 증감에 따라 K-S 수정통계량 역시 기존의 K-S 통계량과 같은 증감을 보이나 K-S 수정통계량은 확률분포의 종류에 따라 차이가 발생한다. 여러 확률분포를 적용시켜도 K-S 통계량은 동일한 결과가 발생하지만 K-S 수정통계량을 적용시키면 분포의 특성을 반영하여 약간씩 차이가 발생한다. 3절 결과와 같이 K-S 통계량과 K-S 수정통계량의 차이는 4%정도로 K-S 통계량과 비교할 때 큰 값이다. K-S 수정통계량을 이용하면 Joseph (2005)가 제안한 K-S 통계량 검정기준에 의한 등급의 결과도 달라질 수 있다. 3.1절에서 언급한 것과 같이 K-S 통계량에서는 'Very Strong'이 오분류율 5%에서만 발생하지만 K-S 수정통계량에서는 5%～20%의 모든 오분류율에서 'Very Strong'한 결과가 나타나기 때문이다.

신용평가 연구에서 사용하는 자료는 분포의 성격을 잘 알고 있으므로 확률밀도함수 $f(\cdot)$ 또는 $F(\cdot)$를 추정 가능하다 가정 하에, 분할된 두 분포함수 $F_D(\cdot)$와 $F_N(\cdot)$의 동일성에 대한 검정방법으로 수정된 K-S 방법이 사용가능하다고 본다. 예제를 통한 결과에서도 확인할 수 있듯이 기존의 방법에 비해 큰 값을 갖는다. 따라서 두 분포함수가 동일하다는 귀무가설을 기각시키기가 더욱 어려워지고 동일한 분포함수라고 확신할 수 있다. 그러므로 기존의 분포를 가정하지 않은 K-S 검정방법보다 자료의 특성이라 할 수 있는 분포를 적용시킨 K-S 수정통계량을 제안하며, 이는 기존의 신용평가에서 사용하는 적합성 검정방법을 개선한 방법이라고 할 수 있겠다.

분포함수가 식 (1.2)와 같이 분할되었고 전체 분포함수를 알고있다는 가정하에 신용평가 연구에서의 자료와 유사한 상황을 설정하여 본 연구에서 제안한 K-S 수정통계량에 관하여 토론하였다. 즉 정규분포와 양의 왜도를 갖는 치우친 분포를 고려하였으며 부도율총합 α를 10%로 선정하여 결과를 살펴보았다. 향후 연구에서는 동일한 가정하에서 연구범위를 확대하여 음의 왜도 또는 첨도 등을 고려한 다양한 확률분포에서 표본을 추출하고 부도율총합의 비율도 다양하게 설정한 K-S 수정통계량에 관한 연구를 제안한다.

참고문헌

송문섭, 박창순, 이정진 (2003). <S-Link를 이용한 비모수 통계학>, 자유아카데미.

Azzalini, A. (1985). A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, **12**, 171–178.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution, *Journal of the Royal Statistical Society, Series B*, **61**, 579–602.

Barton, D. E. and Mallows, C. L. (1965). Some aspects of the random sequence, *Annals of Mathematical Statistics*, **36**, 236–260.

Buccianti, A. (2005). Meaning of the λ parameter of skew-normal and log-skew normal distributions in fluid geochemistry, *CODAWORK'05*, 19–21.

Chang, F. C., Gupta, A. K. and Huang, W. J. (2002). Some skew-symmetric models, *Random Operators and Stochastic Equations*, **10**, 133–140.

Chiogna, M. (1998). Some results on the scalar skew-normal distribution, *Journal of the Italian Statistical Society*, **7**, 1–13.

Daniel, W. W. (1990). *Applied Nonparametric Statistics*, 2nd ed., PWS-KENT, Boston.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramer-von mises tests, *Annals of Mathematical Statistics*, **28**, 823–838.

Genton, M. G. (2005). Discussion of the skew-normal, *Scandinavia Journal of Statistics*, **32**, 189–198.

Gupta, A. K. and Chen, T. (2001). Goodness-of-fit test for the skew-normal distribution, *Communications in Statistics-Simulation and Computation*, **30**, 907–930.

Gupta, A. K., Nguyen, T. and Sanqui, J. A. T. (2004). Characterization of the skew-normal distribution, *Annals of the Institute of Statistical Mathematics*, **56**, 351–360.

Hájek, J., Šidák, Z. and Sen, P. K. (1998). *Theory of Rank Tests*, 2nd ed., Academic Press, New York.

Henze, N. (1986). A probabilistic representation of the skew-normal distribution, *Scandinavian Journal of Statistics*, **13**, 271–275.

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

Liseo, B. (1990). The skew-normal class of densities: Inferential aspects from a Bayesian viewpoint, *Statistica*, **50**, 71–82.

Salvan, A. (1986). Locally most powerful invariant tests of normality, *Atti Della XXXIII Riunione Scientifica Della Societa Italiana di Statistica*, **2**, 173–179.

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples, *Bulletin of Mathematical University of Moscow*, **2**, 3–16.

Tasche, D. (2006). Validation of internal rating systems and PD estimates, *Working paper*, http://arxiv.org/physics/0606071v1

Modified Kolmogorov-Smirnov Statistic for Credit Evaluation

C. S. Hong[1] · G. Bang[2]

[1]Dept. of Statistics, Sungkyunkwan University;
[2]Research Institute of Applied Statistics, Sungkyunkwan University

(Received August 2008; accepted October 2008)

Abstract

For the model validation of credit rating models, Kolmogorov-Smirnov(K-S) statistic has been widely used as a testing method of discriminatory power from the probabilities of default for default and non-default. For the credit rating works, K-S statistics are to test two identical distribution functions which are partitioned from a distribution. In this paper under the assumption that the distribution is known, modified K-S statistic which is formulated by using known distributions is proposed and compared K-S statistic.

Keywords: Credit rating model, score, discriminatory power, distribution function, nonparametric test, probability of default, risk, validation.

[1]Corresponding author: Professor, Dept. of Statistics, Sungkyunkwan University, 3-53, Myungryun-Dong, Jongro-Gu, Seoul 110-745, Korea. E-mail: cshong@skku.ac.kr
[2]Researcher, Research Institute of Applied Statistics, Sungkyunkwan University, 3-53, Jongro-Gu, Seoul 110-745, Korea. E-mail: bgwhite@skku.edu