Journal of the Korean Data & Information Science Society
2013, **24**(1), 85–94

http://dx.doi.org/10.7465/jkdi.2013.24.1.85
한국데이터정보과학회지

목표변수의 형태에 따른 신용평점 모형 구축[†]

우현석[1] · 이석형[2] · 조형준[3]

[123]고려대학교 통계학과

접수 2012년 11월 9일, 수정 2012년 12월 24일, 게재확정 2013년 1월 7일

요 약

금융시장의 규모가 점점 더 커짐에 따라 고객정보 관리 미숙 또는 부실한 의사결정, 즉 신용 리스크 관리 실패로 인한 손실이 막대하게 증가하고 있다. 따라서 신용 리스크 관리가 점차 더 중요해지고, 이런 신용 리스크를 최소화하는 기본적인 도구인 신용 평점 모형이 절실히 요구된다. 신용평점 모형은 주로 이항형 목표변수만 이용하여 개발 연구되었다. 본 논문에서는 순서형 다항 자료 또는 경시적 이항 자료 같은 다른 형태의 목표 변수를 고려한 신용평점 모형구축 방법을 제시한다. 그 개발된 모형을 실제 자료와 랜덤화한 자료에 적용하여 Kolmogorov-Smirnov 통계량으로 비교 분석한다.

주요용어: 경시적 이항 자료, 순서형 다항 자료, 신용 리스크, 신용평점 모형.

1. 머리말

신용 리스크 (credit risk)란 신용과 관련된 사건들로 인해 발생할 수 있는 잠재적인 손실 위험을 의미하는데, 신용등급 (Kim과 Ha, 2010)이나 신용 스프레드 (credit spread)의 변화 등이 여기에 해당되고, 금융 거래에 있어서 고객의 장기적인 대출금 연체나 지급불능 상태와 같은 채무불이행 리스크 (default risk) 또한 이러한 신용 리스크에 해당한다고 할 수 있다 (Bielecki와 Tomasz, 2002). 특히 금융시장의 규모 자체가 커짐에 따라 거래 고객에 대한 불충분한 심사나 고객의 신용 정보 부족으로 인한 손실이 금융 기관에 큰 영향을 주게 되었다. 이러한 원인으로 인해 야기되는 리스크를 줄이기 위해서는 관련 정보들을 기반으로 개별 고객들의 신용도에 대한 정확한 예측과 그에 따른 의사결정 과정이 필요한데, 이를 위해 대부분의 금융기관에서는 그들만의 신용평가시스템 (credit scoring system)을 제작 도입하고 있다. 이 때 개별 고객들의 신용을 예측하고 평점화 (scoring) (Jung, 2010)하는데 있어서 여러 가지 신용평점 모형 (credit scoring model)이 고려되고 활용되는데, 이러한 모형에는 회귀분석 (regression), 신경망 기법 (neural network), 판별분석 (discrimination analysis), 의사결정나무 접근방법 (decision tree approach) 등이 있고, 그 중 평점표 개발에 가장 일반적으로 이용되는 통계적 방법은 로지스틱 회귀모형 (logistic regression model)이다 (Kim, 2004). 이는 예측력과 설명력이 뛰어나고 미국 등 선진국에서 이미 그 효과가 검증되었으며 또한 사람들에게 가장 친숙한 방법이기 때문이다.

일반적으로 신용 평점화 (credit scoring)는 각각의 고객을 두 개의 그룹, 우량과 불량으로 분류하는 문제로 주로 로지스틱 회귀분석 모형을 통해 신용도를 예측하고 알맞은 신용평점을 부여하였다. 또한,

[†] 이 논문은 2012년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2010-0007936).
[1] (136-701) 서울시 성북구 안암동, 고려대학교 정경대학 통계학과, 석사졸업.
[2] (136-701) 서울시 성북구 안암동, 고려대학교 정경대학 통계학과, 석박사통합과정수료.
[3] 교신저자: (136-701) 서울시 성북구 안암동, 고려대학교 정경대학 통계학과, 부교수.
E-mail: hj4cho@korea.ac.kr

정확한 분류를 위해 신용평점에 영향을 크게 미치는 설명변수를 찾는 방법을 연구하였다. 하지만, 목표변수의 형태가 항상 이항형 (binary)인 한계가 존재하였다. 따라서 본 논문에서는 목표변수가 순서형 다항형 또는 경시적 이항형 목표 변수인 경우에 사용할 신용평점 모형 방법을 제시한다.

본 논문의 2절에서는 목표변수 형태에 따른 신용평점 모형에 대한 설명을 할 것이다. 그리고 3절에서는 신용평점 모형에 따른 평점화 방법을 이야기하고, 4절에서는 실제사례에 대한 모형적합 결과를 살펴볼 것이다. 마지막으로 5절에서는 앞의 내용들을 종합하고 결론지어서 마무리할 것이다.

2. 목표변수 형태에 따른 회귀 모형

이 장에서는 신용평점 모형 구축을 위해 사용할 회귀모형을 목표 변수 형태에 따라 정의한다.

2.1. 로지스틱 회귀모형

로지스틱 회귀분석은 목표변수 Y_i는 신용도가 불량이면 0, 우량이면 1인 이항형 목표변수를 모형의 설명변수 $(X_{1i}, \cdots, X_{pi})$들의 비선형 방정식을 통해 설명하고 또 예측하고자 하는 기법으로, 이 때의 모형은 다음과 같다.

$$\log\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi},\ i = 1, \cdots, n \tag{2.1}$$

식 (2.1)은 신용도가 불량인 확률과 우량인 확률의 로그-오즈비를 회귀계수가 $\beta_0, \beta_1, \cdots, \beta_p$인 선형 회귀로 모형화 할 수 있는데, 이는 목표변수인 신용도와 해당 설명변수 사이의 관계에 대한 해석뿐만이 아니라 개별 자료에 대한 추정확률의 계산 또한 가능하여 개별 고객의 신용도를 예측하고 평점화하는데 유용하게 이용될 수 있다.

2.2. 비례오즈모형

비례오즈모형 (proportional odds model)은 목표변수가 세 개 이상의 값을 가지면서 동시에 그 값에 순서가 있을 경우에 분석에 이용되는 기법으로, 목표변수가 단순히 다항일 때에 주로 이용되는 기준 범주 로지스틱 모형 (base-category logistic model)과는 차이가 있다.

$$\log\left(\frac{P(Y_i \le j)}{1-P(Y_i \le j)}\right) = \alpha_j + \beta_1 X_{1i} + \cdots + \beta_p X_{pi},\ i = 1, \cdots, n,\ j = 1, \cdots, J-1 \tag{2.2}$$

식 (2.2)는 목표변수의 누적확률을 이용하고 J-1개의 누적모형에 대해서 설명변수들의 효과가 동일하다고 가정하며 각각의 누적모형의 절편 α_j를 갖는다 (Agresti, 2002). 이 모형에서 회귀계수 $\beta_0, \beta_1, \cdots, \beta_p$를 추정하여 목표변수의 범주에 따른 누적확률을 얻을 수 있고, 각 설명변수의 단위 오즈비 계산도 할 수 있다. 따라서 비례오즈모형은 순서를 갖는 다항의 신용도와 설명변수들 사이의 관계에 대한 분석을 통해 개별 고객의 신용을 예측하고 평가하기 위한 모형으로 적합하다고 할 수 있다.

2.3. Goncalves와 Azzalini 모형

목표변수의 형태가 이항형이지만 시간에 따라 다른 값을 갖는 경시적 이항형 (longitudinal binary)인 자료는 Goncalves와 Azzalini(2008)가 제안한 모형을 사용한다. Y_{it}는 시점 t에서의 신용도 불량을 의

미하고 설명변수 X_{it}도 t시점에서의 신용도 관련 특성 변수로 제안된 모형은 다음과 같다.

$$\log\left(\frac{P(Y_{it}=1)}{1-P(Y_{it}=1)}\right)=\beta_0+\beta_1X_1+\cdots+\beta_pX_{pit},\ i=1,\cdots,n,\ t=1,\cdots,T \tag{2.3}$$

이 방법은 GEE (generalized estimating equation)와 같이 목표변수의 주변 분포 에 대한 모수적 모형을 이용하여 자료를 분석하지만, GEE와는 달리 개별 프로필들에 대한 하나의 완전히 명시된 확률모형을 가지기 때문에 가능도 추론 (likelihood inference)이 가능하다 (Goncalves 등, 2012).

계열 의존도 (serial dependence)를 결정하는 부분에 있어서는 다른 이행 모형 (transition model)과 마찬가지로 마르코프 연쇄 (Markov chain) 방법에 의해 규정되어 진다고 가정하는데, 특이한 점은 로지스틱 모형의 모수들이 그들의 의미를 보존할 수 있도록 적절한 모수화 (parameterization)가 적용된다는 것이다. 이 때의 시차 (lag) 1과 시차 2 의존도 모수 (dependence parameter)들로는 자기 상관 (autocorrelation)보다 오즈비 (odds ratio)를 선호하고 아래와 같은 형태를 갖는다.

$$OR(Y_t,Y_{t-1})=\frac{P\{Y_{t-1}=1,Y_t=1\}P\{Y_{t-1}=0,Y_t=0\}}{P\{Y_{t-1}=0,Y_t=1\}P\{Y_{t-1}=1,Y_t=0\}} \tag{2.4}$$

$$OR(Y_{t-1},Y_{t-2})=\psi_1=OR(Y_{t-1},Y_t) \tag{2.5}$$

$$OR(Y_{t-2},Y_t|Y_{t-1}=0)=\psi_2=OR(Y_{t-2},Y_t|Y_{t-1}=1) \tag{2.6}$$

$$\lambda=(\lambda_1,\lambda_2)=(\log\psi_1,\log\psi_2) \tag{2.7}$$

위의 식을 통해 성공 확률과 의존도가 독립적으로 변화하게 되고, 모수의 해석이 가우스 확률과정 (gaussian process)의 부분 자기상관 (partial autocorrelation)에 대한 것과 비슷해지게 된다 (Goncalves와 Azzalini, 2008).

이 방법을 통해 설명변수와 시간변수 그리고 특정 설명변수와 시간변수 사이의 상호작용 항들에 대한 추정 회귀계수를 얻게 되고, λ를 통해서는 계열 의존도에 대한 두 개의 값을 얻을 수 있는데, 이는 목표변수와 설명변수들 간의 관계를 파악하고 각 시점에서의 성공확률을 추정할 수 있게 한다. 따라서 시간에 따른 변화를 고려하는 모형으로 적합하다고 할 수 있다.

3. 신용평점 모형에 따른 평점화 방법

3.1. 신용 평점표를 이용한 평점화 방법

일반적인 신용평점 모형은 로지스틱 회귀모형을 이용하여 적합한 신용 평점표 (scorecard)를 만든다. 이 과정에서 먼저 각각의 연속형 혹은 이산형 설명변수들은 주로 몇 개의 범주들로 분류 및 재그룹화 되어 분석에 이용되는데, 이를 통해서 몇 가지 장점들을 얻을 수 있다 (Hand와 Adams, 2000). 이렇게 설명변수의 이산화를 통해 만들어지는 신용 평점표는 특정 변수의 특이값에 민감하지 않게 하고 예측력을 높게 만든다. 또한 실제 자료에 흔히 있을 수 있는 각 변수에서의 결측값들을 하나의 그룹으로 처리하기 때문에 자료의 손실 문제를 해결 할 수 있다. 설명변수를 이산화 하는 방법으로는 일반적으로 분위수 (quantile)나 일정 간격 (equal interval)을 이용하는 방법이 있고, 모의 담금질 (simulated annealing) (Hand와 Adams, 2000), 스플라인 분류 기계 (classification spline machine) (Koo 등, 2009) 등이 있다. 신용 평점표를 만드는 과정에 있어서도 다양한 방법을 고려할 수 있다. 그 중 PDO (points to double the odds)는 오즈를 두 배로 만드는 지점의 수치를 이용한 평점표 작성 방법이다. 예를 들어 PDO를 20으로 정했을 경우 이 방법을 통해 520점을 받은 고객은 500점을 받은 고객보다 오즈가 두 배 더 크다

고 해석 할 수 있다. 이러한 PDO 과정은 식 (3.1)을 통해 각 변수들의 범주에 대한 회귀계수 추정치가 양의 값을 가지도록 만든다.

$$\text{보정된 추정치} = \text{회귀계수 추정치} - \text{가장 작은 회귀계수 추정치} \tag{3.1}$$

이후, 적절한 PDO값을 정하여 식 (3.2)와 같이 보정된 회귀계수를 하나의 점수로 선형 변환시킨다 (Kang 등, 2006).

$$\text{점수} = \text{보정된 추정치} \times [PDO/\log 2] \tag{3.2}$$

마지막으로 신용평점 (credit score)의 범주가 0점에서 1000점 사이가 되도록 보정과정을 거치면 신용 평점표가 완성된다. 신용 평점표를 통해, 설명변수에 속한 개인의 값에 따라 신용평점 점수의 증가 여부와, 모든 설명변수들에 대한 신용평점 점수의 함으로 총 신용평점을 계산할 수 있다. 이는 신용평점 점수에 대한 해석 및 이해가 쉬워지는 장점을 갖는다.

3.2. 로지스틱 회귀모형의 평점화 방법

일반적인 평점화 방법과 달리 모형의 추정확률을 이용하여 평점화하는 방법을 제안하였다. 먼저 식 (2.1)을 통해 얻어진 추정확률에 1000을 곱하면 0점에서 1000점 사이의 값을 얻을 수 있는데 이를 개별 데이터에 대한 신용평점으로 사용한다. 이렇게 만들어진 신용평점은 모형을 그대로 이용한다는 점에서 직관적이면서도 자연스러운 방법이라 할 수 있고, 또한 이후의 추가적인 보정과정 없고 평점화 과정에 모형의 절편이 고려된다는 점, 그리고 본래 모형이 가지는 특성과 설명변수와 목표변수 간의 비선형적인 관계를 유지한다는 점에서 기존의 방법과 차이가 있다.

3.3. 비례오즈모형의 평점화 방법

목표변수가 순서형인 경우 비례오즈모형 (proportional odds model)을 통해 얻게 되는 목표변수의 범주에 따른 추정확률을 이용하여 평점화한다. 이는 비례오즈모형이 가지는 목표변수와 설명변수 간의 관계에서의 특성을 그대로 유지할 수 있고 개별 데이터를 평점화 하는데 있어서 각 범주에 따른 정보를 모두 고려할 수 있는 장점이 있다. 하지만 하나의 신용평점은 개개인의 신용이 얼마나 좋은지를 평가하는 척도이기 때문에 모든 정보를 균등하게 고려하기 보다는 좋은 범주일수록 상대적으로 더 높은 가중치를 부여하여 이용하는 것이 더 적절하다고 할 수 있다. 따라서 최종적으로는 각 범주에서의 추정확률의 가중평균을 이용함으로써 개별 데이터의 신용평점을 계산하게 된다. 목표변수의 범주가 $1, 2, \cdots, J$를 가질 때, 신용평점은 다음과 같다.

$$\text{신용평점}_i = 1000 \times \left(\frac{1}{w_{(J)} - w_{(1)}} \right) \times \left[\left(\sum_j w_j \times \widehat{\pi}_{ij} \right) - w_{(1)} \right] \tag{3.3}$$

여기서 $w_j = j/(1 + \cdots + J)$이고, $w_{(1)}, w_{(J)}$는 최소, 최대 가중치를 의미한다. 또한, $\widehat{\pi}_{ij}$는 목표변수 j의 i번째 자료의 추정확률을 의미한다. 위의 식은 일반적인 경우, 목표변수가 가지는 값이 $\{1, 2, \cdots, J\}$와 같은 경우의 것으로, 신용평점은 0점에서 1000점 사이의 범주를 가지게 된다.

3.4. Goncalves와 Azzalini모형의 평점화 방법

목표변수가 경시적 이항형인 자료는 동일한 설명변수를 갖는 이항형 목표변수가 시간의 흐름에 따라 변하는 자료로, 2.1절의 로지스틱 회귀모형을 적용하기에는 시간적 특성을 고려하지 않기 때문에 적절한

방법이라 할 수 없다. 따라서 2.3절의 Goncalves와 Azzalini모형을 사용하여 시간 특성을 반영한 분석결과를 통해 평점화 방법을 부여한다. 경시적 이항형 자료를 한 시점에서만 보면 일반적인 이항형 자료와 형태가 같다. 따라서, 특정 시점에서의 평점화 방법은 분석결과의 추정확률을 이용한 3.1절의 평점화 방법을 사용한다.

$$i\text{번째 자료의 시점 } t\text{에서의 추정 확률 } \widehat{\pi}_{it} = \frac{\exp(x_{it}^T\widehat{\beta})}{1+\exp(x_{it}^T\widehat{\beta})} \tag{3.4}$$

식 (3.4)를 통해 시점에 따른 목표변수의 추정확률을 얻을 수 있고, 이 추정확률에 1000을 곱하여 얻은 신용평점은 0점에서 1000점 사이의 범주를 가지게 된다. 또한, 각 시간별로 똑같은 평점화 방법을 적용하였기 때문에 개별 데이터의 시간에 따른 신용평점 추이를 살펴볼 수 있으므로, 이를 통해 현 시점에서의 정보뿐만이 아니라 향후 일정 시간동안의 정보를 동시에 고려할 수 있고 그럼으로써 좀 더 합리적이고 효과적인 의사결정을 내릴 수 있을 것이다.

4. 사례분석

4.1. 평가방법 및 자료소개

일반적으로 이항형 자료의 신용평점은 평점표를 사용하였다. 본 논문에서는 이와 다르게 로지스틱 모형을 사용하여 평점화를 하는 방법을 설명하였는데, 이 두 방법을 비교 평가하기 위해 누적분포함수와 Kolmogorov-Smirnov (K-S) 통계량을 이용하였다. 여기서 누적분포함수는 신용평점에 따른 누적 우량 및 불량 비율 분포를 나타낸 것으로, 제안한 모형이 적합한 신용평점을 부여한다면 우량고객과 불량고객에 대한 평점의 분포가 잘 분리되고 따라서 그래프 상에서의 두 누적 분포 사이의 거리도 멀어지게 된다. 이 때, s를 모형으로부터 얻은 신용평점이라 한다면, 경험적 확률분포 $P(s)$는 다음과 같다.

$$P(s) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le s) \tag{4.1}$$

G는 우량 신용자, B는 불량 신용자라 한다면, K-S통계량은 다음과 같다.

$$K\text{-}S\text{통계량} = \max_s \{P_G(s) - P_B(s)\} \tag{4.2}$$

식 (4.2)는 우량일 때의 평점 분포와 불량일 때의 평점 분포가 얼마나 멀리 떨어져 있는지를 측정한다 (Thomas 등, 2002). 일반적으로는 신용평점의 범주를 구간화 한 후 각 구간에 따른 누적 백분위 분포표를 이용하여 구한 K-S 통계량 값이 20∼40의 값을 가질 때 평점 모형이 적당하다고 판단하고 40 이상일 경우에는 좋은 모형으로 간주한다 (Kim, 2004). 여기서 두 방법을 비교 분석하기 위해서 SAS의 HMEQ 자료를 사용하였다. 이 자료는 은행에 대출을 신청한 고객자료로 수입 부채비율, 가장 오래된 거래 개월 수, 부실거래수, 현재 자산가치, 주요 부실거래수로 5개의 설명변수와 대출상환여부인 이항형 목표변수로 구성되었다. 목표변수는 4771명의 우량고객과 1189명의 불량고객으로 구성되어 있고, 설명변수의 이산화는 「고객관계관리 (CRM)를 위한 데이터마이닝 방법론」 제 4장에 제시된 방법을 따랐다 (Kang 등, 2006). 목표변수가 이항형 (binary)일 때에는 위와 같이 누적분포함수와 K-S 통계량이라는 기존의 평가 방법이 있었지만 목표변수가 순서형일 경우 적합한 평가척도가 없다. 따라서 본 논문에서는 기존 K-S 통계량의 개념을 기초로 한 새로운 평가척도를 제시하고자 한다. 식 (4.1)을 이용하여 목표변수의 범주가 $i < j$인 경험적 확률을 각각 $P_i(s), P_j(s)$라 하면, 최소 K-S 통계량은 다음과 같다.

$$\text{최소 } K\text{-}S \text{ 통계량} = \min_{i,j}\max_s\{P_i(s) - P_j(s)\} \tag{4.3}$$

최소 K-S 통계량이라고 부르는 이 척도는 한 쌍의 누적 백분위 분포 사이에서 얻게 되는 K-S 통계량들 중 가장 작은 값을 대표값으로 결정하는데, 이는 하나의 신용 자료에 대해서 신용평점 모형이 제대로 작동하고 있는지를 판단하는 잣대가 된다. 다른 쌍에서 얻게 되는 K-S 통계량은 적어도 최소 K-S 통계량보다는 큰 값을 가지게 되고, 따라서 최소 K-S 통계량이 충분히 큰 값을 가진다면 신용평점에 따른 모든 누적 백분위 분포 사이의 거리는 충분히 멀 다고 할 수 있기 때문이다. 목표변수가 순서형일 때 사용된 자료는 Card 자료로 국내 특정 카드회사 고객들의 카드사용 내용 자료로 개인정보를 제외한 당월 일시불 금액, 당월 카드 이용건수 및 전월 카드이용금액, 카드 최초개설일자로부터의 기간과 총 한도 소진율, 5개의 설명변수와 하나의 카드 등급 코드인 순서형 목표변수로 구성되어 있다. 카드 등급은 일반은 1, 우량은 2, Platinum은 3으로 관측되었고 총 83007개의 자료가 관측되었다. 마지막으로 목표변수가 경시적 이항형 (longitudinal binary)도 목표변수가 순서형일 때와 마찬가지로 적합한 평가척도가 없었다. 따라서 본 논문에서는 경시적 이항형인 목표변수에 대해서는 Goncalves와 Azzalini모형의 평점화 방법과 식 (4.3)을 사용하였다. 이 때 사용된 자료는 SAS 및 데이터마이닝에서 제공하는 Buytest자료로 24개월동안 우편(또는 전화)주문으로 60을 초과하여 구매한 고객을 대상으로 10000명을 무작위 추출하여 우편 조사한 자료이다. 6개월 단위의 거래회수 여부를 반응변수로 하여 총 3회의 구매여부를 반응변수로 하였다. 사용된 설명변수는 총 3회의 시간과 연령, 구입총액, 결혼여부, 집의 소유여부, 거주지, 할인고객 여부이다.

4.2. 결과분석

HMEQ자료는 0점에서 1000점 사이의 신용점수를 50점 단위로 구간화 한 누적 백분위 분포표를 사용하였다. 방법1은 PDO를 통한 신용 평점표를 만들어 얻은 결과이고 방법2는 모형의 추정확률을 이용하여 얻은 결과이다. Figure 4.1에서 볼 수 있듯이 방법2가 오히려 기존의 방법1보다 미세하지만 더 좋다는 것을 보여주고 있다. 그리고 Table 4.1은 두 방법을 이용한 평점화 과정을 독립적으로 100번씩 시행한 후 그로부터 구해지는 K-S 통계량의 평균값, 표준오차이다. Figure 4.1과 Table 4.1을 통해서 방법2는 고객들의 신용을 상대평가 할 경우에는 방법1과 다르지 않다는 것이다. 따라서 방법1이 갖는 편리성을 고려하지 않는다면, 방법2를 사용하는데 아무 문제가 없다고 할 수 있으며, 이후에 다른 형태의 목표변수를 가지는 신용자료를 평점화 하는데 방법2를 이용해도 무방할 것이다.

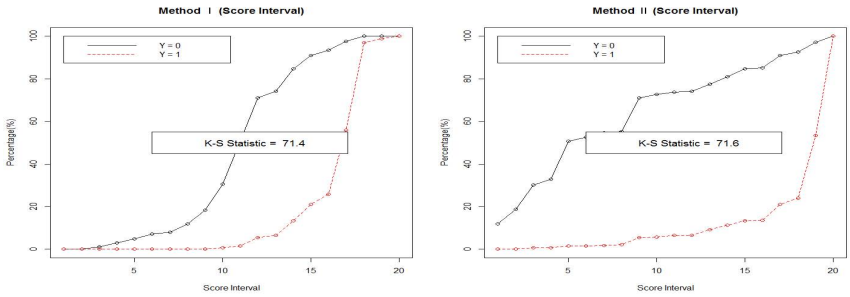

Figure 4.1 Cumulative distribution functions & K-S statistics for HMEQ data

Table 4.1 K-S statistics for HMEQ data

	Method Ⅰ	Method Ⅱ
Avg	67.40	67.78
Std error	0.22	0.20

Card 자료는 목표변수가 순서형 자료로 비례오즈모형을 통해 얻은 신용평점 점수를 50점 단위로 구간화 한 누적 백분표를 사용하였다. 또한 이 방법과 비교하기 위하여 목표변수를 순서를 랜덤하게 바꾸어 생성한 자료를 동일한 방법으로 구한 신용평점 점수를 구간화하여 비교하였다. 만약 제안한 모형이 타당하다면 목표변수를 랜덤하게 순서를 바꾸어 추출한 자료는 잘못된 자료이기 때문에 제안한 모형의 결과가 안 좋게 나타나고 결국 두 방법에 큰 차이가 나타나기 때문이다. Figure 4.2는 평점화 과정을 한 번 시행했을 때 각 자료에서 얻게 되는 누적분포함수와 최소 K-S 통계량을 보여주고 있다. 목표변수가 가지는 값들에 상관없이 신용평점에 따른 누적 백분위 분포들이 잘 분리되어 있고 최소 K-S 통계량 또한 만족할만한 수준이기 때문에, 이를 통해 목표변수가 순서형일 때 제안된 평점화 방법이 적절하다는 것을 알 수 있다. 또한 Table 4.2를 통해 동일 한 방법을 100번 반복했을 때 구해지는 평균과 표준 오차의 값을 통하여 실제자료가 목표변수를 임의화 한 자료보다 좋은 것을 알 수 있으며 이는 새로운 평가 방법을 뒷받침하는 결과라 할 수 있다.

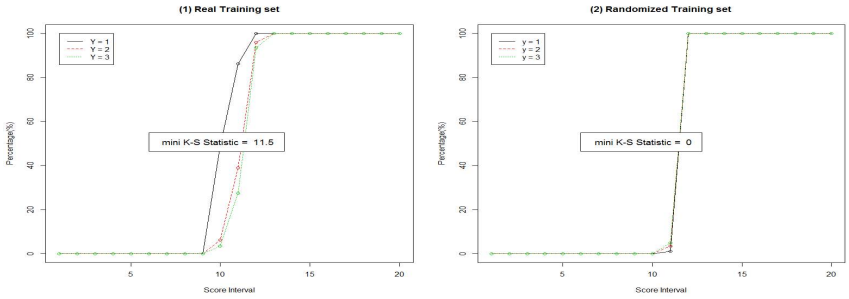

Figure 4.2 Cumulative distribution functions & K-S statistics for card data

Table 4.2 K-S statistics for card data

	Real Data	Random Data
Avg	12.18	0.20
Std error	0.07	0.06

다음으로는 목표변수가 경시적 이항형일 때를 살펴보았다. 본래 자료와 목표변수를 순서를 랜덤하게 바꾸어 생성한 자료를 Goncalves와 Azzalini의 모형을 통해 얻은 신용 평점 점수를 비교하였다. Figure 4.3과 이를 100번 반복한 Table 4.3을 통해 실제자료가 목표변수의 순서를 랜덤하게 추출한 자료보다 모든 시점에서 잘 분류하는 것을 알 수 있다. 또한, 이는 경시적 이항형인 자료의 신용 평점화에 있어 적절한 척도가 됨을 뒷받침하는 결과라 할 수 있다.

Table 4.3 K-S statistics for buytest data

Time	Real Data		Random Data	
	Avg	Std error	Avg	Std error
T = 1	52.12	0.22	1.23	0.23
T = 2	50.15	0.23	0.88	0.17
T = 3	52.85	0.18	1.07	0.19

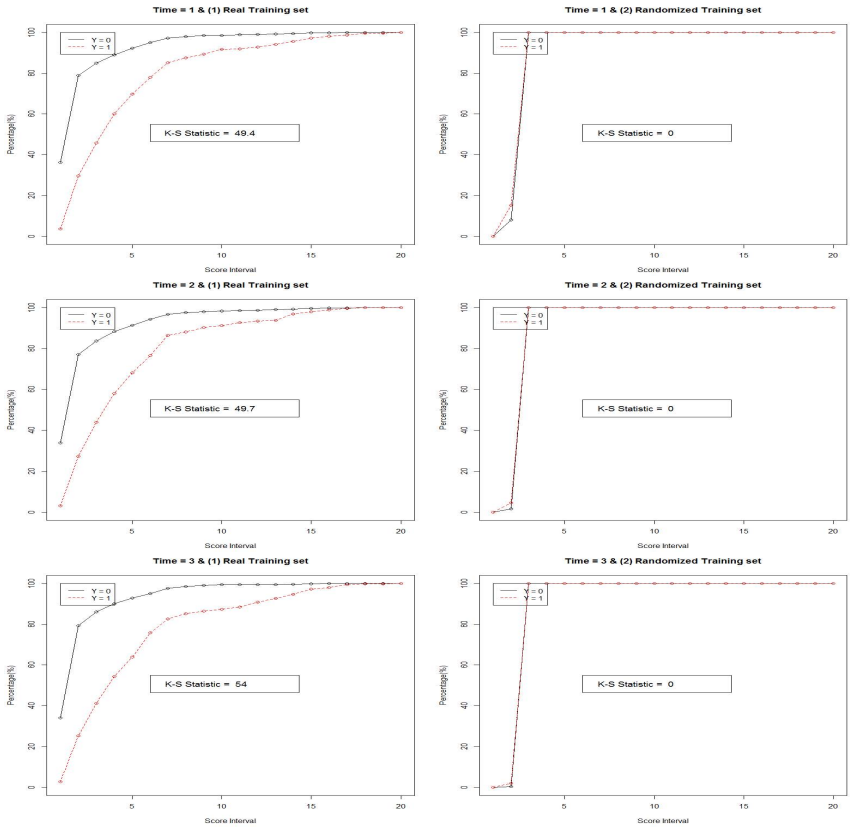

Figure 4.3 Cumulative distribution functions & K-S statistics for buytest data

5. 결론

지금까지 목표변수의 형태에 따른 신용평점 모형을 검토하고 또 제안해 보았다. 실제 자료들을 이용하여 그에 따른 평점화 결과들을 보여주었고 이를 통해 제시된 방법들의 적합성을 살펴보았다.

목표변수가 이항형인 경우에는 기존의 방법과 새로운 평점화 방법을 함께 살펴보았다. 새롭게 제안된 추정확률을 이용한 평점화 방법은 로지스틱 회귀모형을 있는 그대로 이용한다는 점에서 직관적이고 자연스러운 방법이라 할 수 있고, 정확성에서도 기존의 방법과 큰 차이가 없음을 알 수 있었다. 순서형 자료의 경우에는 새로운 하나의 평점 모형을 제안하였다. 비례오즈모형을 이용하여 자료를 분석하였고 그로부터 얻어지는 추정 확률들의 가중평균의 형태를 하나의 평점화 방법으로 채택함으로써 자료를 평점화 하였다. 경시적 이항형 목표변수인 경우에는 Goncalves와 Azzalini (2008)가 제안한 모형을 통해 자료를 분석해 보았다. 시작시점에서의 특성을 기반으로 하여 시점에 따른 추정확률 값들을 얻을 수 있었고, 이를 이용한 평점화 방법을 각 시점마다 적용함으로써 자료를 평점화 하였다. 평점화 방법에 대한 평가척도로는 기존 K-S 통계량의 확장된 개념이라고 할 수 있는 최소 K-S 통계량을 이용하였다. 각 자료에서 두 가지의 형태의 훈련자료를 활용하였는데, 원자료와 목표변수를 무작위로 추출하여 생성한 자료를 통해 모형의 평가척도의 적합성을 검증하였다. 목표변수 형태에 따라 제안한 모형이 적절하다면,

설명변수와 목표변수간에 연광성이 있는 자료에서는 좋은 평가척도가 나타나야 하고, 그렇지 않으면 나쁜 평가척도가 나타나야 한다. 따라서 원자료로부터 얻어진 평가척도와 목표변수를 무작위로 추출하여 설명변수와 연관성이 없는 자료에서 얻어진 평가척도를 비교하여 제안한 모형의 타당성을 보였다.

금융시장의 규모가 점차적으로 확대됨에 따라 신용 리스크 또한 증대되고 있고, 이에 각각의 금융기관들은 예측 가능한 손실에 대한 위험을 줄이기 위해 고객에 대한 정보관리 및 의사결정 과정 개발에 몰두하고 있다. 지금까지 신용평점 모형에 대한 개발 및 보완은 이러한 맥락에서 비롯되었다고 할 수 있다. 본 논문에서 다뤘던 내용 또한 그 동안 다루지 않았던 목표변수를 대상으로 신용평점 모형을 고려했다는 것에 의의가 있다. 또한 제시한 방법들을 여러 자료들에 실제로 적용시켜봄으로써 활용 가능성을 검증받을 수 있다면, 좀 더 효율적이고 정확한 신용평점 모형 구축 할 수 있을 것이라 생각한다.

참고문헌

Agresti, A. (2002). *Categorical data analysis*, 2nd Ed., Wiley-Interscience, New York.

Bielecki, T. R. and Rutkowski, M. (2002). *Credit risk: Modeling, valuation and hedging*, Springer-Verlag, Berlin.

Gonsalves, M. H. and Azzalini, A. (2008). Using markov chains for marginal modelling of binary longitudinal data in an exact likelihood approach. *Metron*, **2**, 157-181.

Gonsalves, M. H., Cabral, M. S. and Azzalini, A. (2012). The R package bild for the analysis of binary longitudinal data. *Journal of Statistical Software*, **9**, 1-17.

Hand, D. J. and Adams, N. M. (2000). Defining attributes for scorecard construction in credit scoring. *Journal of Applied Statistics*, **27**, 527-540.

Jung, K. M. (2010). Development of educational software for coarse classifying and model evaluation in credit scoring. *Journal of the Korean Data & Information Science Study*, **21**, 1225-1235.

Kang, H. C., Han, S. T., Choi, J. H., Lee, S. G., Kim, E. S., Um, I. H. and Kim, M. K. (2006), *Methodology of data mining for C.R.M. : A case study on Enterprise Miner*, Free Academy, Seoul.

Kim, E. N. and Ha, J. (2010). Study on the validation methods of calibration considering correlations. *Journal of the Korean Data & Information Science Study*, **21**, 407-417.

Kim, M. J. (2004). *Understanding and applying credit scores*, ePharos, Seoul.

Koo, J., Park, C. and Jhun, M. (2009). A classification spline machine for building a credit scorecard. *Journal of Statistical Computation and Simulation*, **79**, 681-689.

Thomas, L. C., Edelman, D. B. and Crook, J. L. (2002). *Credit scoring and its applications*, SIAM, Philadelphia.

Journal of the Korean Data & Information Science Society
2013, **24**(1), 85–94

http://dx.doi.org/10.7465/jkdi.2013.24.1.85
한국데이터정보과학회지

Building credit scoring models with various types of target variables[†]

Hyun Seok Woo[1] · Seok Hyung Lee[2] · HyungJun Cho[3]

[123]Department of Statistics, Korea University

Received 9 November 2012, revised 24 December 2012, accepted 7 January 2013

Abstract

As the financial market becomes larger, the loss increases due to the failure of the credit risk managements from the poor management of the customer information or poor decision-making. Thus, the credit risk management also becomes more important and it is essential to develop a credit scoring model, which is a fundamental tool used to minimize the credit risk. Credit scoring models have been studied and developed only for binary target variables. In this paper, we consider other types of target variables such as ordinal multinomial data or longitudinal binary data and suggest credit scoring models. We then apply our developed models to real data and random data, and investigate their performance through Kolmogorov-Smirnov statistic.

Keywords: Credit risk, credit risk management, credit scoring model, longitudinal binary data, ordinal multinomial data.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2010-0007936).

[1] Master of Science, Department of Statistics, Korea University, Seoul 136-701, Korea.
[2] Ph.D. student, Department of Statistics, Korea University, Seoul 136-701, Korea.
[3] Corresponding author: Associate professor, Department of Statistics, Korea University, Seoul 136-701, Korea. E-mail: hj4cho@korea.ac.kr