*Journal of the Korean Data & Information Science Society*
*2011,* **22**(2), 149–157

한국데이터정보과학회지

# 신용평가에서 로지스틱회귀를 이용한 미결정자 추론

홍종선[1] · 정민섭[2]

[1]성균관대학교 통계학과 · [2]성균관대학교 응용통계연구소

접수 2011년 1월 3일, 수정 2011년 1월 24일, 게재확정 2011년 2월 1일

**요 약**

본 연구는 신용평가 과정에서 발생하는 미결정자를 결측자료 문제로 간주하여 MAR와 MNAR 가정 하에서 추론한다. MAR 가정에서 미결정자 추론은 결정자들에 대한 로지스틱 회귀모형의 회귀계수벡터를 이용하여 미결정자의 부도 확률을 구한 후 결정자의 부도확률과 비교하여 미결정자의 미래 상태를 판단한다. 그리고 MNAR 가정에서의 미결정자 추론은 특성변수가 추가한 로지스틱모형으로부터 미결정자의 부도확률을 구하고 미결정자를 예측하는 방법을 제안하였다. 두 종류의 실제 자료에 대하여 모의실험을 한 결과, MAR 가정에서 미결정자의 비율이 증가하더라도 원자료의 오분류율과 추론한 결과 차이가 없으며, MNAR 가정에서는 추가적인 변수를 고려하여 미결정자를 추정하였기 때문에 미결정자의 오분류율이 MAR 가정에서의 오분류율보다 감소하고 나아가 전체에서 미결정자가 차지하는 비율이 증가함에 따라 전체의 오분류율이 더욱 감소함을 발견하였다.

주요용어: 결측자료, 로지스틱모형, 부도율, 오분류율, 평점, 혼동행렬.

## 1. 서론

신용평가 (credit evaluation) 또는 신용평점 (credit scoring) 제도는 구축된 차주 (borrower)의 자료를 바탕으로 신용심사와 관리업무에 사용한다. 신용의 정도를 평점으로 제시하기 위하여 신용상태의 자료를 분석하여 평가모형을 설정하고 평점을 추정한다. 대출자 (은행)는 차주의 평점에 근거하여 정상 (우량; good, non-default)인 차주에게는 대출해주고 부도 (불량; bad, default)를 예상한 차주에게는 대출을 억제하면서 거래를 차등적으로 적용하여 대출자의 수익을 증대하고 위험을 최소화하는 목적이 있다 (Kim과 Lee, 2003; 홍종선과 김지훈, 2009; 홍종선과 권태완, 2010).

신용평가 과정에서 미결정 (undecided) 차주가 발생하는 문제가 발생하는데 다음과 같이 두 종류의 원인으로 구분될 수 있다. 첫 번째는 여러 가지 이유로 심사가 아직 이루어지지 않고 판단이 보류되어 미결정 차주가 발생하며, 두 번째로는 신용평가를 판단하기 어려운 평점 때문에 평가를 유보하고 특별한 전문가에게 재심사를 의뢰하기 위하여 결정이 보류된 미결정 차주가 발생하는 경우이다.

미결정 차주가 미래에 부도 또는 정상 차주인지를 예측하는 것은 매우 힘들고 어렵기 때문에 결정된 차주만으로 모형을 개발하고, 결정된 차주와 미결정된 차주가 모두 포함되어 있는 차주 집단에 대해 적용한다면 큰 편의 (bias)가 발생한다. 이런 문제점을 해결하기 위하여 본 연구에서는 미결정된 차주를 추론하는 방법을 연구한다.

미결정 추론 문제는 신용평가에서 차주의 신용상태를 예측하는 문제 뿐만 아니라, 여러 다른 분야에서 추론문제로 활용할 수 있다. 예를 들어 의학에서 의사가 환자의 질병유무의 판단을 내리기 어려운 경우

[1] 교신저자: (110-745) 서울 종로구 명륜동 3-53, 성균관대학교 통계학과, 교수. E-mail: cshong@skku.ac.kr
[2] (110-745) 서울 종로구 명륜동 3-53, 성균관대학교 응용통계연구소 연구원, 일반대학원 통계학과 석사과정.

에는 판단을 보류하여 전문의에게 심사를 의뢰하거나 추가적인 검사로 얻은 결과로 다시 판단한다. 또한 테니스, 펜싱, 크릿켓 (Ananda, 2010) 등의 여러 종목의 운동경기에서 선수가 심판의 판단에 불복하여 재심사를 요구하면 다른 심판관이 심사하거나 컴퓨터를 이용한 정밀 판독으로 심사한다.

본 연구에서는 신용평가 분야에서의 결정 또는 미결정 차주를 일반적인 분야에서 활용할 수 있도록 결정자 또는 미결정자로 설정하면서 설명하고 상세한 상황 설명이 필요한 부분에서만 신용평가 용어로 설명한다. 신용평가 용어를 사용하여 예를 들어 설명하면, 평점이 낮으면 재무상태를 포함한 신용상태가 좋지 않으므로 대출자는 부도가 날 것으로 예측하여 대출자가 대출을 허락하기 어려운 상황이며, 평점이 높으면 신용상태가 매우 좋다고 판단하여 대출을 허락해주는 상황을 의미한다.

Feelders (2000)와 Hand (2001)는 미결정자 추론을 결측자료 (missing data) 문제로 간주하였는데 Little과 Rubin (1987), Feelders (2000)와 Kim (2002)이 정의한 결측값 유형을 살펴보면서 미결정자의 종류를 2절에서 정의한다. 3절에서는 두 종류로 분류된 미결정자의 추론과정을 각 종류별로 제안하고, 4절에서는 실증예제를 통해 두 종류의 미결정자 추론방법의 결과를 얻고 원자료와 비교분석한다. 마지막 5절에서 결론을 유도한다.

## 2. 미결정자 정의

관찰된 차주의 신용정보 자료를 확률벡터 $\boldsymbol{X}=(X_1,\ldots,X_k)$로 표기한다. 차주의 신용상태를 정상과 부도로 구분하는 확률변수 $Y$를 다음과 같이 나타내며,

$$Y=\begin{cases}1 & \text{만약 부도인 경우}\\ 0 & \text{만약 정상인 경우.}\end{cases}$$

대출여부를 결정한 결정자인지 또는 미결정자인지를 알려주는 보조변수 $A$를 다음과 같이 정의한다.

$$A=\begin{cases}1 & \text{만약 거절자인 경우}\\ 0 & \text{만약 승인자인 경우}\\ u & \text{만약 미결정자인 경우.}\end{cases}$$

미결정자는 결측자료 문제로 간주하는데 결측값의 유형은 MCAR (missing completely at random), MAR (missing at random) 그리고 MNAR (missing not at random)로 구분한다 (Little과 Rubin, 1987; Feelders, 2000). 심사가 보류되어 발생된 미결정자는 미결정자를 제외한 결정자와 동일한 속성을 갖고 있어서 부도로 판단할 확률이 동일하기 때문에 Feelders (2000)의 거절자 추론문제와 동일하게 간주하여 MAR 방법으로 접근하며, 전문가에게 재심사를 의뢰하기 위하여 결정이 보류된 미결정자 추론문제는 독립변수의 조건부 미결정자는 신용상태에 의존한다. 이 경우의 결정자의 분포는 미결정자의 분포와 다르기 때문에 MNAR 방법으로 간주한 새로운 방법을 제안한다. 본 연구에서는 미결정자 문제를 MAR과 MNAR의 결측자료 방법으로 구분하여 추론한다.

### 2.1. MAR 가정의 미결정자

신용평가모형으로부터 얻은 $i$번째 차주의 신용평점을 확률변수 $S_i$라고 하고, 절단점 (승인점; cutoff point, threshold) $c$를 기준으로 $i$번째 차주의 미래상태를 부도 또는 정상 차주로 예측하여 판단하고, 이와 관계없이 결정이 보류된 차주를 미결정자라 정의하고 아래와 같이 나눌 수 있다.

$$A_1=\begin{cases}1 & \text{만약 } S_i\le c\\ 0 & \text{만약 } S_i>c\\ u & \text{만약 미결정자인 경우.}\end{cases} \tag{2.1}$$

심사가 보류되어 발생된 미결정자 그룹 ($A_1 = u$)은 미결정자를 제외한 결정자그룹 ($A_1 \neq u$)과 동일한 속성을 갖고 있으며 부도로 판단할 확률이 동일하기 때문에 Feelders (2000)는 거절자 추론 (reject inference) 문제로 간주하여 MAR 방법으로 접근하였다. 이와 동일하게 본 연구에서는 MAR 가정에서의 미결정자를 새롭게 표현하면 다음과 같다. $\boldsymbol{X}$=$\boldsymbol{x}$=$(x_1, \ldots, x_k)$의 조건부 미결정자는 $Y$에 의존하지 않기 때문에

$$P(A_1 = u|\boldsymbol{X} = \boldsymbol{x}, Y = y) = P(A_1 = u|\boldsymbol{X} = \boldsymbol{x}).$$

이며 다시 정리하면 다음과 같다 (Feelders, 2000).

$$P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) \ = P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}, A_1 = u) = P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}, A_1 \neq u),$$

여기서 $A_1 \neq u = A_1 = 01$.

### 2.2. MNAR 가정의 미결정자

두 개의 절단점 $c_1, c_2$에 대하여 $i$번째 차주의 미래상태를 부도와 정상 차주로 판단하며, 두 절단점 사이에 해당하는 차주의 신용평가는 판단하기 어렵기 때문에 평가를 유보하여 전문가에게 재심사를 의뢰하기 위하여 결정이 보류된 미결정자로 판단하는 보조변수 $A_2$는 다음과 같이 정의한다.

$$A_2 = \begin{cases} 1 & \text{만약 } S_i \leq c_1 \\ u & \text{만약 } c_1 < S_i \leq c_2 \\ 0 & \text{만약 } S_i > c_2. \end{cases} \tag{2.2}$$

전문가에게 재심사를 의뢰하기 위하여 결정이 보류된 미결정자 추론문제에서는 $\boldsymbol{X} = \boldsymbol{x}$의 조건부 미결정자가 $Y$에 의존하기 때문에

$$P(A_2 = u \mid \boldsymbol{X} = \boldsymbol{x}, Y = y) \neq P(A_2 = u \mid \boldsymbol{X} = \boldsymbol{x})$$

이며, 결정자 $Y$의 분포는 미결정자 $Y$의 분포와 다르기 때문에 MNAR 방법으로 간주할 수 있다. 즉

$$P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}, A_2 = u) \neq P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}, A_2 \neq u).$$

위의 두 확률은 $P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$와 동일하지 않다. 그리고 이 경우의 미결정가 추론은 결정자로부터 유도한 모형과 다른 모형을 사용해야 한다.

## 3. 미결정자 추론 과정

### 3.1. MAR 가정의 미결정자 추론

미결정자들 ($A_1 = u$)과 미결정자를 제외한 결정자들 ($A_1 \neq u$)이 동일한 속성을 갖기 때문에, 우선 결정자 ($A_1 \neq u$)들에 대한 로지스틱 회귀모형 (logistic regression model)을 설정한다.

$$\ln \frac{\hat{P}}{1 - \hat{P}} = b_0 + \boldsymbol{b}'\boldsymbol{X}, \tag{3.1}$$

여기서 $\hat{P} = \hat{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}, A_1 = u)$이다. 로지스틱모형 (3.1)으로부터 얻은 특성변수의 회귀계수 벡터 ($b_0$, $\boldsymbol{b}$)를 사용하여 미결정자들의 부도확률을 다음과 같이 예측할 수 있다.

$$\hat{P} = \frac{e^{b_0+\boldsymbol{b}'\boldsymbol{x}}}{1+e^{b_0+\boldsymbol{b}'\boldsymbol{x}}}. \tag{3.2}$$

최종적인 신용평가는 미결정자 $(A_1 = u)$의 $\boldsymbol{X}$=$\boldsymbol{x}$를 사용하여 추정한 부도확률 식 (3.2)를 이용하여 다음과 같이 판단한다.

$$\hat{Y} = \begin{cases} 1 & \text{만약} \quad \hat{P} \le P^C \\ 0 & \text{만약} \quad \hat{P} > P^C, \end{cases} \tag{3.3}$$

여기서 평가기준인 $P^C$은 결정자 $(A_1 \neq u)$들 중에서 부도로 판단된 차주의 확률로 다음과 같이 정의한다.

$$P^C = P(Y = 1 \mid A_1 \neq u). \tag{3.4}$$

그러므로 MAR 가정 하의 미결정자 추론 절차는 다음과 같다.

1. 결정자들에 대한 로지스틱 회귀모형 (3.1)으로부터 특성변수 $\boldsymbol{X} = (X_1, \ldots, X_k)$에 대한 회귀계수 벡터 $(b_0,\ \boldsymbol{b})$를 구한다.
2. 결정자들에 대한 추정된 회귀계수벡터를 이용하여 (3.2)식의 미결정자의 부도확률 $\hat{P}$을 계산한다.
3. 미결정자의 부도확률 $\hat{P}$와 (3.4)식의 평가기준 $P^C$와 비교하여 미결정자의 미래상태를 (3.3)의 함수를 이용하여 추정한다.

### 3.2. MNAR 가정의 미결정자 추론

미결정자 $Y$의 분포는 결정자의 분포와 다르기 때문에 미결정자 추론은 결정자로부터의 모형과 다른 모형을 사용해야한다. 따라서 결정자의 신용평가모형에 사용한 변수 $\boldsymbol{X} = (X_1, \ldots, X_k)$외에 미결정자의 신용평가에 도움을 주는 확률변수 $\boldsymbol{X}^+ = (X_{k+1}, \ldots . X_p)$을 추가한 특성변수 $\boldsymbol{X}^* = (\boldsymbol{X}, \boldsymbol{X}^+) = (X_1, \ldots, X_k, X_{k+1}, \ldots, X_p)$를 사용하여 다음과 같은 로지스틱 회귀분석을 한다.

$$\ln(\frac{\hat{P^*}}{1-\hat{P^*}}) = b_0 + \boldsymbol{b}'\boldsymbol{X}^*, \tag{3.5}$$

여기서 $\hat{P^*} = \hat{P}(Y = 1 \mid \boldsymbol{X}^* = \boldsymbol{x}^*, A_2 = u)$은 특성변수가 $\boldsymbol{X}^* = \boldsymbol{x}^*$일 때의 미결정자의 부도확률로 다음과 같이 구한다.

$$\hat{P^*} = \frac{e^{b_0+\boldsymbol{b}'\boldsymbol{x}^*}}{1+e^{b_0+\boldsymbol{b}'\boldsymbol{x}^*}}. \tag{3.6}$$

$\boldsymbol{X}^* = \boldsymbol{x}^*$를 사용하여 추정된 미결정자의 부도확률로 다음과 같이 신용평가를 예측한다.

$$\hat{Y} = \begin{cases} 1 & \text{만약} \quad \hat{P^*} \le P^C \\ 0 & \text{만약} \quad \hat{P^*} > P^C. \end{cases} \tag{3.7}$$

MNAR 가정 하의 미결정자 추론은 다음과 같은 절차를 따른다.

1. 결정자의 신용평가모형에 사용한 특성변수에 추가된 특성변수 $\boldsymbol{X}^*$=$\boldsymbol{x}^*$를 사용하여 로지스틱모형 (3.5)로부터 (3.6)식의 미결정자 부도확률 $\hat{P^*}$을 구한다.
2. 미결정자의 부도확률 $\hat{P^*}$과 (3.4)식의 평가기준 $P^C$와 비교하여 미결정자의 미래상태를 (3.7)의 함수를 이용하여 추정한다.

## 4. 실증예제

### 4.1. 실증예제1

신용평가분야의 예제는 4.2절에서 분석하고 우선 일반적인 분야인 의학예제를 먼저 들어본다. Pepe (1998, 2003)에서 분석한 자료를 실증예제로 채택하고 표 4.1과 같이 청각장애의 유무에 관한 자료로 반응변수로 청각장애(Y) 그리고 특성(독립)변수로 가청범위($X_s$), 소리강도($X_l$), 소리주파수($X_f$), 청력한계($X_{amt}$)를 고려한다.

**표 4.1** 실증예제1 자료구성

| 범주 | 변수이름 | 변수설명 | 형태 |
|---|---|---|---|
| 반응변수 | $Y$ | 청각장애 | 범주형(0,1) |
| 특성변수 | $X_s$ | 가청범위 | 연속형 |
| | $X_l$ | 소리강도 | 연속형 |
| | $X_f$ | 소리주파수 | 연속형 |
| | $X_{amt}$ | 청력한계 | 연속형 |

식 (2.1)에서의 절단점 $c$는 특성변수 $\boldsymbol{X} = (X_s, X_l, X_f)$를 사용하여 얻은 전체자료에서의 부도율 $P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = 0.2646$으로 설정하고, 이 절단점 c를 기준으로 작성된 2×2 혼동행렬 (confusion matrix)는 표 4.2과 같다.

**표 4.2** 원자료의 혼동행렬 ($N$=1848)

| | | 실제 | |
|---|---|---|---|
| | | 장애 | 정상 |
| 결정 | 장애 | 410 | 185 |
| | 정상 | 79 | 1,174 |

MAR 가정에서 미결정자는 전체표본의 10%를 단순 무작위 추출 (simple random sampling)하여 $N = 185$인 집단으로 선정한다. 미결정자의 추론은 3.1절에서 제안한 방법을 사용한다. 즉 각각의 미결정자는 (3.4)식의 평가기준 $P^C$와 비교하여 미결정자의 미래상태를 추정하여 얻은 결과를 표 4.3 왼쪽에 정리하였다. 전체 자료 중에서 미결정자로 선정된 표본 자료의 추론 결과와 나머지 자료를 통합하여 얻은 혼동행렬은 표 4.3 오른쪽에 나타내었다.

**표 4.3** MAR 가정의 미결정자 추론결과와 추론 후 전체의 혼동행렬

| 미결정자 추론 | | 실제 | |
|---|---|---|---|
| | | 장애 | 정상 |
| 결정 | 장애 | 39 | 22 |
| | 정상 | 7 | 117 |

| 전체 | | 실제 | |
|---|---|---|---|
| | | 장애 | 정상 |
| 결정 | 장애 | 410 | 189 |
| | 정상 | 79 | 1,170 |

그리고 MNAR 가정에서 미결정자는 3.2절에서 제안한 추론방법을 사용한다. MAR 가정에서 전체자료의 부도율을 절단점으로 얻은 c=0.2646 를 기준으로 전체의 상위 5%와 하위 5%를 각각 식 (2.2)의 $c_1$과 $c_2$로 설정하여 $c_1$과 $c_2$사이에 존재하는 전체자료의 10%를 MNAR 가정에서의 미결정자로 설정한다. 미결정자의 추론은 $X_{amt}$가 추가된 특성변수 $\boldsymbol{X}^* = (X_s, X_l, X_f, X_{amt})$를 사용하여 식 (3.5)의 로지스틱 회귀모형으로 추정된 부도확률을 얻어 (3.7)의 관계식으로부터 신용평가를 추정하여 얻은 결과를 표 4.4 왼쪽에 정리하였다. 그리고 전체 자료 중에서 MNAR 가정에 미결정자로 선정된 표본 자료의 추론 결과와 나머지 자료를 통합하여 생성한 혼동행렬은 표 4.4 오른쪽에 나타내었다.

**표 4.4** MNAR 가정의 미결정자 추론결과와 추론 후 전체의 혼동행렬

| 미결정자 추론 | | 실제 | |
|---|---|---|---|
| | | 장애 | 정상 |
| 결정 | 장애 | 49 | 0 |
| | 정상 | 1 | 132 |

| 전체 | | 실제 | |
|---|---|---|---|
| | | 장애 | 정상 |
| 결정 | 장애 | 430 | 122 |
| | 정상 | 59 | 1237 |

본 연구에서 제안한 두 종류의 추론 방법이 MAR과 MNAR 가정에서 이루어지고 있기 때문에 두 종류를 비교하기 위하여 미결정자의 비율을 전체 자료에서 6% (상하위 3%씩)와 14% (상하위 7%씩)로 동일하게 설정하고 분석한 결과는 표 4.5와 같다.

**표 4.5** 실증예제1에서 미결정자 추론 결과 비교 (실제 오분류율 (14.28%))

| 가정 | MAR | | | MNAR | | |
|---|---|---|---|---|---|---|
| 미결정자 비율 | 6% | 10% | 14% | 6% | 10% | 14% |
| 오분류율(%) | 14.33 | 14.50 | 14.17 | 11.53 | 9.79 | 7.91 |
| 실제와 차이(%) | -0.05 | -0.22 | 0.11 | 2.75 | 4.49 | 6.37 |

각각의 MAR과 MNAR 가정에서 미결정자의 비율을 6%, 10%, 14%로 바꿔가며 실험하여 표 4.5에 결과를 정리하였다. 표 4.5를 살펴보면 MAR 가정에서 미결정자의 비율이 증가하여도 원자료의 오분류율과 추론한 결과 차이가 거의 변화가 없었다. 이것은 MAR 가정에서 결정자와 미결정자의 부도와 정상의 특성이 같다고 가정하였기 때문에 결정자와 미결정자의 특성이 같아 미결정자의 추정결과가 결정자와 비슷하게 나왔음을 알 수 있다. MNAR 가정의 추정결과에서는 결정자와 미결정자의 부도와 정상의 특성이 다르다고 가정하였기 때문에 미결정자를 잘 설명할 수 있는 특성변수를 추가하여 미결정자를 추정함에 따라 오분류율이 감소됨을 알 수 있고, 미결정자 비율이 증가하면서 전체적으로 오분류율이 줄어든다는 것을 파악할 수 있다. 이 결과로부터 기존의 특성변수 외에 미결정자를 잘 추정하는 특성변수의 정보가 추가되면서 정분류율이 증가하고 오분류율이 감소하는 효과가 발생했다고 해석할 수 있다.

### 4.2. 실증예제2

1994년부터 2005년까지 외감 기업 중 총 자산규모가 4,500억 이상인 한국 기업에 대한 자료이다 (홍종선과 최진수, 2009). 총 4,134개의 자료와 119개의 변수 중 네개의 변수를 선택하였다. 자료의 구성은 아래 표 4.6과 같다.

**표 4.6** 실증예제2 자료구성

| 범주 | 변수이름 | 변수설명 | 형태 |
|---|---|---|---|
| 반응변수 | $Y$ | 부도여부 | 범주형(0,1) |
| 특성변수 | $X_s$ | 기존점수 | 연속형 |
| | $X_a$ | 차입금 의존도 | 연속형 |
| | $X_b$ | 자산대비 부채비율 | 연속형 |
| | $X_c$ | 금융비용비율 | 연속형 |

특성변수 $\boldsymbol{X} = (X_s, X_a, X_b)$를 사용하여 얻은 전체 자료의 부도율 $P = P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$=0.0576을 식 (2.1)의 절단점 c를 기준으로 작성된 $2 \times 2$ 혼동행렬는 표 4.7과 같다.

MAR 가정에서 미결정자는 표본크기의 10%를 단순 무작위 추출하여 N=413인 미결자 집단으로 선정하여 3.1절에서 제안한 미결정자의 추론방법을 사용한다. 미결정자의 미래상태를 추정하여 얻은 결과

**표 4.7** 원자료의 혼동행렬 ($N$=4134)

| | | 실제 | |
|---|---|---|---|
| | | 부도 | 정상 |
| 결정 | 거절 | 204 | 675 |
| | 승인 | 34 | 3,221 |

를 표 4.8 왼쪽에 정리하였다. 그리고 전체 자료 중에서 미결정자로 선정된 표본 자료의 추론 결과와 나머지 자료를 통합하여 얻은 혼동행렬은 표 4.8 오른쪽과 같다.

**표 4.8** MAR 가정의 미결정자 추론결과와 추론 후 전체의 혼동행렬

| 미결정자 추론 | | 실제 | |
|---|---|---|---|
| | | 부도 | 정상 |
| 결정 | 거절 | 25 | 68 |
| | 승인 | 4 | 316 |

| 전체 | | 실제 | |
|---|---|---|---|
| | | 부도 | 정상 |
| 결정 | 거절 | 204 | 660 |
| | 승인 | 34 | 3,236 |

MNAR 가정에서 미결정자는 $X_c$변수가 추가된 특성변수 $\boldsymbol{X}^* = (X_s, X_a, X_b, X_c)$를 사용하여 식 (3.5)의 로지스틱 회귀모형으로 추정된 부도확률을 얻어 개개인의 신용평가를 예측하여 얻은 결과를 표 4.9 왼쪽에 정리하였다. 그리고 전체 자료 중에서 MNAR 가정에서 미결정자로 선정된 표본 자료의 추론 결과와 나머지 자료를 통합하여 생성한 혼동행렬은 표 4.9 오른쪽에 나타내었다. 미결정자의 비율을 전체 자료에서 6%와 14%로 바꿔가며 분석한 결과는 표 4.10과 같다.

**표 4.9** MNAR 가정의 미결정자 추론결과와 추론 후 전체의 혼동행렬

| 미결정자 추론 | | 실제 | |
|---|---|---|---|
| | | 부도 | 정상 |
| 결정 | 거절 | 29 | 1 |
| | 승인 | 2 | 375 |

| 전체 | | 실제 | |
|---|---|---|---|
| | | 부도 | 정상 |
| 결정 | 거절 | 219 | 471 |
| | 승인 | 19 | 3,425 |

**표 4.10** 실증예제2에서 미결정자 추론 결과 비교 (실제 오분류율 (17.15%))

| 가정 | MAR | | | MNAR | | |
|---|---|---|---|---|---|---|
| 미결정자 비율 | 6% | 10% | 14% | 6% | 10% | 14% |
| 오분류율(%) | 16.79 | 16.81 | 16.59 | 13.74 | 11.85 | 9.94 |
| 실제와 차이(%) | 0.36 | 0.34 | 0.56 | 3.41 | 5.70 | 7.21 |

실증예제2의 결과를 종합적으로 정리한 표 4.10과 실증예제1의 결과 표 4.5와 비교하면 유사한 경향을 나타내고 있음을 확인할 수 있다. 따라서 MAR 가정에서는 미결정자와 결정자의 부도와 정상의 특성이 같다고 가정하였기 때문에 미결정자의 비율과 상관없이 오분류율이 실제 오분류율과 근사함을 보였다. MNAR 가정의 추정결과에서는 결정자와 미결정자의 특성이 다르다고 가정하였기 때문에 미결정자의 오분류율이 MAR 가정에서의 오분류율보다 감소하고 나아가 전체에서 미결정자가 차지하는 비율이 증가함에 따라 전체의 오분류율이 더욱 감소한다는 사실을 확인한다. 그러므로 미결정자의 신용상태를 잘 설명하는 특성변수의 정보가 추가되면서 정분류율이 증가하고 오분류율이 감소하는 효과가 발생한다고 판단할 수 있다.

## 5. 결론

운동경기에서 선수가 심판의 판단에 불복하여 재심사를 요구하면 미결정된 판단이라고 간주하여 이를 다른 심판관이 심사하거나 컴퓨터를 이용한 정밀 판독으로 심사하기도 하며, 의학에서도 의사가 환자의 질병유무의 판단을 내리기 어려운 경우 판단을 보류하여 미결정된 판단을 전문의사에게 의뢰하기도 하거나 추가적인 검사로 얻은 결과로 다시 판단하는데, 본 연구는 신용평가 과정에서 발생하는 미결정자의 종류를 두가지로 분류하고 각각의 미결정자를 추론하는 방법을 제안하였다.

미결정자 추론은 결측자료 문제로 간주하여 첫 번째 방법은 심사가 이루어지지 않고 판단이 보류되어 발생하는 미결정자를 MAR 가정에서 추론하고, 두 번째로는 신용평가를 판단하기 어려운 평점 때문에 평가를 유보하고 특별한 전문가에게 재심사를 의뢰하기 위하여 결정이 보류되어 발생하는 미결정자를 MNAR 가정에서 추론한다. MAR 가정에 미결정자 추론은 특성변수가 포함된 결정자들에 대한 로지스틱 회귀모형의 회귀계수벡터를 이용하여 미결정자의 부도확률을 구한다. 미결정자의 부도확률을 평가기준으로 설정한 결정자의 부도확률과 비교하여 미결정자의 미래상태를 판단한다. 그리고 MNAR 가정에서 미결정자 추론은 결정자의 신용평가모형에 사용한 특성변수에 다른 특성변수를 추가한 로지스틱모형으로부터 미결정자의 부도확률을 구하고, 평가기준과 비교하여 미결정자의 미래상태를 추정하는 방법을 제안하였다.

두 종류의 실제 자료에 대하여 미결정자의 비율을 다양하게 설정하고, MAR과 MNAR 가정에서 동일하게 설정하여 비교분석하였다. MAR 가정에서 미결정자의 비율이 증가하더라도 원자료의 오분류율과 추론한 결과 차이가 없었다. 이것은 MAR 가정에서 결정자와 미결정자의 부도와 정상의 특성이 같다고 가정하였기 때문에 결정자와 미결정자의 특성이 같아 미결정자의 추정결과 또한 결정자와 근사함을 파악할 수 있었다. MNAR 가정의 추정결과에서는 결정자와 미결정자의 부도와 정상의 특성이 다르다고 가정하여 추가적인 변수를 고려하여 미결정자를 추정하였기 때문에 미결정자의 오분류율이 MAR 가정에서의 오분류율보다 감소하고 나아가 전체에서 미결정자가 차지하는 비율이 증가함에 따라 전체의 오분류율이 더욱 감소한다는 사실을 확인할 수 있었다. 그러므로 미결정자의 신용상태를 잘 설명하는 특성변수의 정보가 추가되면서 정분류율이 증가하고 오분류율이 감소하는 효과를 발견하였다.

## 참고문헌


홍종선, 권태완 (2010). 수익률 분포의 적합과 리스크값 추정. <한국데이터정보과학회지>, **21**, 219-229.

홍종선, 김지훈 (2009). 신용평가모형에서 두 분포함수의 동일성 검정을 위한 비모수적인 검정방법. <한국데이터정보과학회지>, **20**, 261-272.

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

Ananda, B. W. (2010). Receiver operating characteristic curves for measuring the quality of decisions in cricket. *Journal of Quantitative Analysis in Sports*, **6**, Article 8.

Feelders, A. J. (2000). Credit scoring and reject inference with mixture models. *International Journal of Intelligent System in Accounting*, **8**, 271-279.

Hand, D. J. (2001). Reject inference in credit operations. *Handbook of Credit Scoring*, 225-240.

Kim, H. J. (2002). Analysis of incomplete data with nonignorable missing values. *Journal of the Korean Data & Information Science Society*, **13**, 167-174.

Kim, K. S. and Lee, C. S. (2003). A study of data mining optimization model for the credit evaluation. *Journal of the Korean Data & Information Science Society*, **14**, 825-836.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*, Wiley, New York.

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.

*Journal of the Korean Data & Information Science Society*
*2011,* **22**(2), 149–157

한국데이터정보과학회지

# Undecided inference using logistic regression for credit evaluation

Chong Sun Hong[1] · Min Sub Jung[2]

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

Received 3 January 2011, revised 24 January 2011, accepted 1 February 2011

**Abstract**

Undecided inference could be regarded as a missing data problem such as MAR and MNAR. Under the assumption of MAR, undecided inference make use of logistic regression model. The probability of default for the undecided group is obtained with regression coefficient vectors for the decided group and compare with the probability of default for the decided group. And under the assumption of MNAR, undecided inference make use of logistic regression model with additional feature random vector. Simulation results based on two kinds of real data are obtained and compared. It is found that the misclassification rates are not much different from the rate of raw data under the assumption of MAR. However the misclassification rates under the assumption of MNAR are less than those under the assumption of MAR, and as the ratio of the undecided group is increasing, the misclassification rates is decreasing.

*Keywords*: Confusion matrix, logistic model, misclassification rate, missing data, probability of default.

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