Journal of the Korean Data & Information Science Society
2011, **22**(6), 1017–1028

한국데이터정보과학회지

이변량 프로빗모형을 이용한 미결정자 추론

홍종선[1] · 정미향[2]

[1]성균관대학교 통계학과 · [2]성균관대학교 응용통계연구소

접수 2011년 9월 14일, 수정 2011년 10월 18일, 게재확정 2011년 10월 23일

요 약

신용평가를 판단하기 어렵기 때문에 평가를 유보하고 특별한 전문가에게 재심사를 의뢰하기 위하여 결정이 보류된 미결정자에 대한 미결정자 추론은 신용평가 분야 이외에도 의학통계와 스포츠통계 등 대부분의 통계적 모형에서 발생하는 문제이다. 본 연구에서는 미결정자 추론을 비임의결측 가정하에서의 결측자료 유형으로 간주하고, 표본선택모형 중의 하나인 이변량 프로빗모형을 이용한다. 결정된 차주의 특성을 나타내는 확률변수를 사용하여 미결정자를 추론하는 방법과 보다 정확한 정보를 수집한 후 추가적인 확률변수를 사용하여 추론하는 방법을 제안한다. 실증예제를 통하여 특성변수의 조합과 다양한 미결정 구간, 그리고 절단점의 변동에 따라 미결정자와 전체 오분류율을 비교한다. 미결정구간을 확대하거나 정확한 신용정보를 모형에 추가하여 사용하면 정상 집단과 부도 집단의 정보를 더욱 정확하게 반영할 수 있기 때문에 미결정자와 전체 오분류율의 큰 감소효과를 기대할 수 있다.

주요용어: 거절자추론, 결측자료, 신용평가, 표본선택, 프로빗모형.

1. 서론

신용평점 (credit scoring system) 제도는 고객 중심의 데이터베이스를 구축하여 신용 심사업무 및 고객 관리업무에 사용함으로써 신용상태의 지속성을 가정하여 과거자료를 분석한 평점모형을 추정하고 신용의 정도를 평점으로 제시하며, 이러한 평점에 근거한 업무 프로세스에 따라 거래를 차등적으로 적용, 불량채권의 발생을 사전에 예방함으로써 수익을 증대하며 위험을 최소화한다 (김형준, 2005). 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 최소화한다 (Kim과 Lee, 2003; 홍종선과 정민섭, 2011).

여신을 실행하는 금융회사에서 개인의 신용을 평가할 때 여러가지 이유로 결정을 하지 않거나 보류하여 미결정된 차주가 빈번하게 발생한다. 따라서 미결정자 (undecided)를 제외한 결정자만으로 모형을 개발한다면, 결정자와 미결정자가 모두 포함되어 있는 고객 집단에 대해 적용함에 있어 큰 편의 (bias)를 야기하는 문제에 직면하게 된다. 미결정자의 우 · 불량을 예측하는 작업은 힘들고 조심스럽지만, 미결정자들을 추론하여 신용평가모형에 포함시키는 작업은 중요하다.

Feelders (2000)와 Hand (2001)는 미결정자 추론을 결측자료 (missing data) 문제로 간주하였고 결측값의 유형에 따라 미결정자를 구분하였다. 미결정자 추론은 신용평가 분야에 국한된 문제가 아니라 의학통계와 스포츠통계 등 대부분의 통계적 모형의 분석 과정에서 빈번히 발생하는 문제이다. 예를 들

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

어 의학에서 환자의 질병유무의 판단을 내리기 어려운 경우에 의사가 판단을 보류하여 전문의에게 심사를 의뢰하거나 추가적인 검사로 얻은 결과로 다시 진단하여 판단하며, 테니스, 펜싱, 크리켓 (Ananda, 2010)등의 여러종목의 운동경기에서 선수가 심판의 판단에 불복하여 재심사를 요구하면 다른 심판관이 심사하거나 컴퓨터를 이용한 정밀 판독으로 심사한다.

본 논문에서는 표본선택모형 (sample selection model) 중의 하나인 이변량 프로빗모형 (bivariate probit model)을 이용한 미결정자 추론모형을 연구한다. 미결정자를 포함한 모형을 구현함으로써 결정자만으로 모형을 개발했을 경우에 발생할 수 있는 편의를 줄이고 기존의 미결정자를 추론하는데 시간과 비용을 감소하고, 보다 정확하고 객관적인 판단을 도와주는 하나의 통계방법을 제안한다.

2절에서는 미결정자를 정의하며 미결정자를 결측자료로 간주하여 결측값의 유형들과의 관계를 설명하고, 어느 특정한 결측자료의 유형의 가정에서의 미결정자 추론의 방법인 이변량 프로빗모형을 소개한다. 3절에서는 이변량 프로빗모형을 이용한 미결정자 추론방법을 제안한다. 4절에서는 실증예제를 통해 3절에서 설명한 미결정자 추론방법을 적용하여 분석하고, 마지막 5절에서 결론을 유도한다.

2. 미결정자와 이변량 프로빗모형

2.1. 미결정자

관찰된 차주의 신용정보 자료를 k차원 확률변수의 행렬 $\boldsymbol{X}=(X_1,\ldots,X_k)$로 표기한다. 신용정보 자료를 활용하여 차주의 신용상태를 부도와 정상으로 구분하는 확률변수 Y를 다음과 같이 나타내며

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

대출여부를 결정한 미결정자인지 결정자인지를 알려주는 보조변수 A를 다음과 같이 정의한다:

$$A = \begin{cases} 1 & \text{미결정자인 경우} \\ 0 & \text{결정자인 경우.} \end{cases}$$

미결정자는 결측자료 문제로 간주하는데 결측값의 유형은 완전임의결측 (missing completely at random; MCAR), 임의결측 (Missing At Random; MAR), 비임의결측 (missing not at random; MNAR)로 구분한다 (Little과 Rubin, 1987; Feelders, 2000; Kim, 2002). 미결정자는 크게 두 가지 종류로 존재할 수 있다. 하나는 여러가지 이유로 심사가 아직 이루어지지 않고 판단과 결정이 보류된 경우이고, 다른 하나는 신용평가를 판단하기 어려운 여러가지 상황으로 인하여 평가를 유보하여 전문가에게 재심사를 의뢰하거나 추가적인 평가를 위하여 결정이 보류된 경우이다. 첫 번째 경우의 미결정자는 결측값의 유형 중에서 MAR로 간주하며, 두 번째 경우에는 MNAR로 간주하는데 이에 대하여는 2.2절에서 상세히 설명한다.

2.2. MAR과 MNAR 가정의 미결정자

심사가 보류되어 발생된 미결정자는 미결정자를 제외한 결정자와 동일한 속성을 갖고 있어서 부도로 판단할 확률이 동일하기 때문에 Feelders (2000)의 거절자 추론 (reject inference)문제와 동일하게 간주하여 MAR 방법으로 접근한다. $\boldsymbol{X}=\boldsymbol{x}$의 조건부 결정자는 Y에 의존하지 않기 때문에

$$P(A=0|\boldsymbol{X},Y=y) = P(A=0|\boldsymbol{X})$$

이며, 결정자 Y의 분포는 미결정자 Y의 분포와 같기 때문에 다음과 같이 나타낼 수 있다:

$$P(Y=0|\boldsymbol{X}) = P(Y=0|\boldsymbol{X}, A=1) = P(Y=0|\boldsymbol{X}, A=0).$$

여러가지 복잡한 상황으로 인하여 신용평가를 판단하기 어려운 경우에 전문가에게 재심사를 의뢰하기 위해 판단을 보류하는 경우와 신용평가에 관하여 더욱 정확한 정보를 추가적으로 수집한 후 재심사를 하기위한 과정에서의 미결정자의 분포는 결정자의 분포와 다르며 이 경우에는 MNAR 방법으로 간주할 수 있다. 전문가에게 재심사를 의뢰하기 위하여 결정이 보류된 미결정자 추론문제에서는 $\boldsymbol{X} = x$의 조건부 결정자가 Y에 의존하기 때문에

$$P(A=0|\boldsymbol{X}, Y=y) \neq P(A=0|\boldsymbol{X})$$

이며, 결정자의 분포는 미결정자의 분포와 다르기 때문이다. 즉

$$P(Y=0|\boldsymbol{X}, A=1) \neq P(Y=0|\boldsymbol{X}, A=0).$$

위의 두 확률은 $P(Y=0|\boldsymbol{X})$와 동일하지 않다. 그리고 이 경우의 미결정자 추론은 결정자로부터 유도한 모형과 다른 모형을 사용해야 한다.

MAR 가정은 동일한 속성 하에서 결정자와 미결정자가 동일한 확률을 갖지만 미결정자 그룹 ($A=1$)의 승인 · 거절 결정이 결정자 그룹 ($A=0$)의 특성변수 $\boldsymbol{X}$의 영향력이 다르거나 $\boldsymbol{X}$이외에 추가적으로 영향을 주는 요소가 존재하는 경우가 대부분이다. 동일한 속성을 갖는 경우라도, 미결정자의 우량 확률과 미결정을 제외한 집단의 우량 확률은 다르게 계산되므로 MNAR 가정이 필요하므로 본 연구에서는 MNAR 가정 하에 미결정자를 추론한다.

본 논문에서는 전체 차주 (표본)의 수를 n, 미결정된 차주의 수를 n_1으로 한다 (결정된 차주의 수는 $n-n_1$). 신용평가모형으로부터 얻은 차주의 신용평점을 확률변수 S라고 하면, 미결정자를 다음과 같이 설정할 수 있다. 두 개의 절단점 c_1, c_2 사이 이외의 평점에서는 차주의 미래 상태를 부도와 정상 차주로 확실하게 판단할 수 있으나, 두 절단점 사이의 차주의 평점에서는 신용평가를 판단하기 어렵기 때문에 결정이 보류된 미결정자로 간주한다. 즉 $A = I(c_1 < S \leq c_2)$.

2.3. 이변량 프로빗모형

표본선택모형은 표본선택 편의로부터 발생하는 문제를 해결하기 위해 이용된다 (Heckman, 1979 ; Sartori, 2003). 표본선택모형 중 하나인 이변량 프로빗모형은 2차 이상 일련의 선택에 따른 표본의 선택성을 통제하는데 사용되는 통계적 방법이다 (Greene, 1996; Poirier, 2002). 확률행렬 $\boldsymbol{X}_{1i}$와 $\boldsymbol{X}_{2i}$는 각각 선택한 i번째 관측값에 대한 특성변수를 나타낸다. ε_{1i}과 ε_{2i}는 관찰되지 않는 오차항이다. 오차항 ε_{1i}과 ε_{1i}는 상관관계가 있고, 두 오차항은 이변량 정규분포함수를 나타낸다고 가정한다. 이변량 프로빗모형에서의 선택모형 (selection model)은 다음과 같다.

$$A_i^* = \boldsymbol{X}_{1i}\beta_1 + \varepsilon_{1i}, \quad i = 1, \ldots, n, \tag{2.1}$$

여기서 ε_{1i}는 표준정규분포를 따르고, 선택변수 A_i는 $A_i = I(A_i^* \geq 0)$으로 설정한다. 그리고 이변량 프로빗모형에서의 결과모형 (outcome model)은 다음과 같다.

$$Y_i^* = \boldsymbol{X}_{2i}\beta_2 + \varepsilon_{2i}, \quad i = 1, \ldots, n-n_1, \tag{2.2}$$

여기서 ε_{2i}도 표준정규분포를 따르고, 결과변수 Y_i는 $Y_i = I\left(Y_i^* \geq 0\right)$로 정의한다. 오차항들은 다음과 같은 정규분포를 가정한다.

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{pmatrix}\right), \quad i = 1, \ldots, n - n_1. \tag{2.3}$$

이변량 프로빗모형의 식 (2.1)과 (2.2)은 각각 다음의 의미를 갖는다.

$$\Phi^{-1}\left[P\left(A_i = 0 | \boldsymbol{X}_{1i}\right)\right] = \boldsymbol{X}_{1i}\beta_1 + \varepsilon_{1i}, i = 1, \ldots, n,$$
$$\Phi^{-1}\left[P\left(Y_i = 1 | \boldsymbol{X}_{2i}, A_i = 0\right)\right] = \boldsymbol{X}_{2i}\beta_2 + \varepsilon_{2i}, i = 1, \ldots, n - n_1.$$

차주의 신용상태가 결정된 $A_i = 0$인 경우에는 두 종류의 결과값 Y_i이 관찰되므로 이에 대한 확률을 구할 수 있으며, 미결정된 $A_i = 1$인 경우의 확률은 다음과 같이 구한다 (Feelder, 2000a).

$$\begin{aligned}
P(A_i = 0, Y_i = 0) &= P(A^*{}_i < 0, Y^*{}_i < 0) = P(\boldsymbol{X}_{1i}\beta_1 + \varepsilon_{1i} < 0, \boldsymbol{X}_{2i}\beta_2 + \varepsilon_{2i} < 0) \\
&= P(\varepsilon_{1i} < -\boldsymbol{X}_{1i}\beta_1, \varepsilon_{2i} < -\boldsymbol{X}_{2i}\beta_2) = \Phi_2(-\boldsymbol{X}_{1i}\beta_1, -\boldsymbol{X}_{2i}\beta_2, \rho_2), \\
P(A_i = 0, Y_i = 1) &= P(A^*{}_i < 0, Y^*{}_i \geq 0) = P(\boldsymbol{X}_{1i}\beta_1 + \varepsilon_{1i} < 0, \boldsymbol{X}_{2i}\beta_2 + \varepsilon_{2i} \geq 0) \\
&= P(\varepsilon_{1i} < -\boldsymbol{X}_{1i}\beta_1, \varepsilon_{2i} \geq -\boldsymbol{X}_{2i}\beta_2) = \Phi_2(-\boldsymbol{X}_{1i}\beta_1, \boldsymbol{X}_{2i}\beta_2, -\rho_2),\ i = 1, \ldots, n - n_1, \\
P(A_i = 1) &= P(A_i^* \geq 0) = P(\boldsymbol{X}_{1i}\beta_1 + \varepsilon_{1i} \geq 0) = P(\varepsilon_{1i} \geq -\boldsymbol{X}_{1i}\beta_1) = \Phi(\boldsymbol{X}_{1i}\beta_1),\ i = 1, \ldots, n.
\end{aligned} \tag{2.4}$$

식 (2.4)에서 상관계수 ρ_2가 0이면 다음이 성립하며

$$P\left(Y_i = 0 | \boldsymbol{X}_{2i}, A_i = 1\right) = P\left(Y_i = 0 | \boldsymbol{X}_{2i}, A_i = 0\right),$$

그리고 상관계수 ρ_2가 0이 아닌 값이면 다음과 같이 부등호의 관계로 표현되고

$$\rho_2 > 0\text{일 때}, P\left(Y_i = 0 | \boldsymbol{X}_{2i}, A_i = 1\right) > P\left(Y_i = 0 | \boldsymbol{X}_{2i}, A_i = 0\right),$$
$$\rho_2 < 0\text{일 때}, P\left(Y_i = 0 | \boldsymbol{X}_{2i}, A_i = 1\right) < P\left(Y_i = 0 | \boldsymbol{X}_{2i}, A_i = 0\right).$$

이 확률은 $P\left(Y_i = 0 | \boldsymbol{X}_{2i}\right)$과 다르므로 MNAR 가정을 만족함을 확인할 수 있다.

식 (2.4)에서 나타난 세 종류의 확률을 바탕으로 로그가능도함수를 구성하면 다음과 같고, 이를 최대화하는 최대가능도추정량 (MLE)을 구하여 미결정자 추론에 이용한다.

$$\sum_{A_i=0, Y_i=0} \ln \Phi_2\left(-\boldsymbol{X}_{1i}\beta_1, -\boldsymbol{X}_{2i}\beta_2, \rho_2\right) + \sum_{A_i=0, Y_i=1} \ln \Phi_2\left(-\boldsymbol{X}_{1i}\beta_1, \boldsymbol{X}_{2i}\beta_2, -\rho_2\right) + \sum_{A_i=1} \ln \Phi\left(\boldsymbol{X}_{1i}\beta_1\right). \tag{2.5}$$

SAS의 QLIM (Qualitative and LImited dependent variable Model) 프로시저는 종속변수가 이산형이거나 제한된 범위에서만 관찰이 될 때, 로짓 (logit), 프로빗 (probit), 선택 (selection), 그리고 다변량 (multivariate) 모형을 분석한다. 본 연구에서는 PROC QLIM 프로시저 중 이변량 프로빗모형을 사용하여 미결정자를 추론한다.

3. 미결정자 추론모형

3.1. 특성변수를 이용한 미결정자 추론모형

미결정자 추론을 위한 이변량 프로빗모형을 2절에서와 같이 가정하여 미결정자 추론에 이용한다. 식 (2.5)의 로그가능도함수에 대한 최대가능도추정량 $\hat{\beta}_1, \hat{\beta}_2, \hat{\rho}$ 를 구하고 특성변수 $\boldsymbol{X}_{1i}$, $\boldsymbol{X}_{2i}$를 사용하여 모형 (2.1)과 (2.2)을 다음과 같이 추정한다.

$$\widehat{A}_i^* = \boldsymbol{X}_{1i}\hat{\beta}_1, \quad i = 1, \ldots, n. \tag{3.1}$$

$$\widehat{Y}_i^* = \boldsymbol{X}_{2i}\hat{\beta}_2, \quad i = 1, \ldots, n - n_1. \tag{3.2}$$

식 (3.1)과 (3.2)는 $\Phi^{-1}\left[\widehat{P}\left(A_i = 0 | \boldsymbol{X}_{1i}\right)\right] = \boldsymbol{X}_{1i}\hat{\beta}_1$과 $\Phi^{-1}\left[\widehat{P}\left(Y_i = 1 | \boldsymbol{X}_{2i}, A_i = 0\right)\right] = \boldsymbol{X}_{2i}\hat{\beta}_2$로 표현된다. 따라서 결정된 차주의 신용상태가 부도로 예측하는 경우는 다음과 같이 결정자에 대한 결과모형을 설정한다. 임의의 $i = 1, \ldots, n - n_1$에 대하여,

$$\widehat{Y}_i = \begin{cases} 1, & \widehat{Y}_i^* \geq 0 \text{ 또는 } \Phi\left(\boldsymbol{X}_{2i}\hat{\beta}_2\right) \geq 0.5 \\ 0, & \text{그 외.} \end{cases} \tag{3.3}$$

결정된 차주의 수 $n - n_1$에 대한 결정자 모형을 설정한 과정에서 식 (2.4) 중의 하나인 미결정자의 정보가 포함된 $P(A_i = 1)$와 이를 포함한 식 (2.5)의 로그가능도함수를 이용하였기 때문에 결정자 모형을 바탕으로 평가가 유보된 n_1개의 미결정된 차주를 재평가하여 추론한다. 결정자에 대한 결과모형으로 추정한 식 (3.1), (3.2)를 사용하여 미결정자를 추론한다. 미결정자 $j = 1, \ldots, n_1$ 에 대하여,

$$\widehat{Y}_j = \begin{cases} 1, & \widehat{P}\left(Y_j = 1 | \boldsymbol{X}_{2j}, A_j = 1\right) = \Phi\left(\boldsymbol{X}_{2j}\hat{\beta}_2\right) \geq 0.5 \\ 0, & \widehat{P}\left(Y_j = 1 | \boldsymbol{X}_{2j}, A_j = 1\right) = \Phi\left(\boldsymbol{X}_{2j}\hat{\beta}_2\right) < 0.5. \end{cases} \tag{3.4}$$

3.2. 새로운 특성변수를 이용한 미결정자 추론모형

미결정자의 신용평가는 판단하기 어렵기 때문에 더욱 정확한 정보를 추가적으로 수집한 후 재심사를 의뢰하여 미결정자 추론모형을 설정하고자 한다. 전체 차주의 신용 정보를 나타내는 $\boldsymbol{X}_2$에 추가적인 신용정보로 수집된 $\boldsymbol{X}^+$가 추가된 특성변수 $\boldsymbol{X}_3 = (\boldsymbol{X}_2, \boldsymbol{X}^+)$를 사용하여 다음과 같이 결정자에 대한 이변량 프로빗모형의 결과모형을 가정한다.

$$Y_i^* = \boldsymbol{X}_{3i}\beta_3 + \varepsilon_{3i}, \quad i = 1, \ldots, n - n_1, \tag{3.5}$$

여기서 ε_{3i}의 분포는 표준정규분포로 가정하며 기존의 오차항들의 분포를 종합적으로 설명하면 다음과 같다.

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_2 & \rho_3 \\ \rho_2 & 1 & 0 \\ \rho_3 & 0 & 1 \end{pmatrix}\right)$$

식 (3.4)로부터 결과변수 Y_i가 부도로 판단되는 조건부 확률의 프로빗은 $\Phi^{-1}\left[P\left(Y_i = 1 | \boldsymbol{X}_{3i}, A_i = 0\right)\right] = \boldsymbol{X}_{3i}\beta_3 + \varepsilon_{3i}$으로 표현된다. 식 (2.4)와 같이 차주의 신용상태가 결정된 $A_i = 0$인 경우에는 두 종류의 결과값 Y_i이 관찰되는 확률 $P(A_i = 0, Y_i = 0) = \Phi_2(-\boldsymbol{X}_{1i}\beta_1, -\boldsymbol{X}_{3i}\beta_3, \rho_3)$, $P(A_i = 0, Y_i = 1) = \Phi_2(-\boldsymbol{X}_{1i}\beta_1, \boldsymbol{X}_{3i}\beta_3, -\rho_3)$ 과 미결정된 $A_i = 1$인 경우의 확률은 $P(A_i = 1) = \Phi(\boldsymbol{X}_{1i}\beta_1)$이므로 이를 바탕으로 로그가능도함수를 구성하면 다음과 같다.

$$\sum_{A_i=0, Y_i=0} \ln \Phi_2\left(-\boldsymbol{X}_{1i}\beta_1, -\boldsymbol{X}_{3i}\beta_3, \rho_3\right) + \sum_{A_i=0, Y_i=1} \ln \Phi_2\left(-\boldsymbol{X}_{1i}\beta_1, \boldsymbol{X}_{3i}\beta_3, -\rho_3\right) + \sum_{A_i=1} \ln \Phi\left(\boldsymbol{X}_{1i}\beta_1\right). \tag{3.6}$$

최대가능도추정량 $\hat{\beta}_1, \hat{\beta}_3, \hat{\rho}_3$를 구하고 특성변수 $\boldsymbol{X}_{1i}, \boldsymbol{X}_{3i}$를 사용하여 모형 (2.1)은 모형 (3.1)로 추정하고, 모형 (3.5)를 모형 (3.7)과 같이 추정한다.

$$\widehat{Y}_i^* = \boldsymbol{X}_{3i}\hat{\beta}_3, \quad i = 1, \ldots, n - n_1. \tag{3.7}$$

식 (3.7)은 $\Phi^{-1}\left[\widehat{P}\left(Y_i = 1 | \boldsymbol{X}_{3i}, A_i = 0\right)\right] = \boldsymbol{X}_{3i}\hat{\beta}_3$로 표현되므로 결정된 차주의 신용상태를 부도로 예측하는 경우의 결정자 모형은 다음과 같다. 임의의 $i = 1, \ldots, n - n_1$에 대하여,

$$\widehat{Y}_i = \begin{cases} 1, & \widehat{Y}_i^* \geq 0 \text{ 또는 } \Phi\left(\boldsymbol{X}_{3i}\hat{\beta}_3\right) \geq 0.5 \\ 0, & \text{그 외.} \end{cases} \tag{3.8}$$

3.1절에서 제안한 방법보다 정확한 미결정자 추론을 위하여 차주의 신용에 관한 정보를 추가적으로 수집하여 $\boldsymbol{X}_2$를 포함한 특성변수 $\boldsymbol{X}_3$를 사용하여 식 (3.8)과 같이 $n - n_1$의 결정자 모형을 설정했다. 이 과정에서 미결정자 정보도 포함된 확률 $P(A_i = 1)$와 이를 포함한 로그가능도함수 식 (3.6)을 이용하였기 때문에 결정자 모형을 바탕으로 평가가 유보된 n_1개의 미결정된 차주를 재평가하여 미결정자의 추정값 $\widehat{Y}_j$은 식 (3.8)을 사용하여 다음과 같이 추론한다. 미결정자 $j = 1, \ldots, n_1$에 대하여,

$$\widehat{Y}_j = \begin{cases} 1, & \widehat{P}\left(Y_j = 1 | \boldsymbol{X}_{3j}, A_j = 1\right) = \Phi\left(\boldsymbol{X}_{3j}\hat{\beta}_3\right) \geq 0.5, \\ 0, & \widehat{P}\left(Y_j = 1 | \boldsymbol{X}_{3j}, A_j = 1\right) = \Phi\left(\boldsymbol{X}_{3j}\hat{\beta}_3\right) < 0.5. \end{cases} \tag{3.9}$$

4. 실증예제

1994년부터 2005년까지 외감기업 중 총 자산규모가 4,500억 이상인 한국 기업으로 총 4,134 기업에 대한 자료이다 (홍종선과 최진수, 2009). 총 $n = 4,134$개의 자료와 119개의 변수로 구성되어있는데 단계적 변수선택 방법을 사용하여 부도여부를 판단하는데 결정적인 역할은 하는 네 개의 독립변수를 선택하였고 실제 부도가 발생했는지를 나타내는 신용상태의 결과를 나타내는 변수와 같이 고려하였다. 전체 기업중 부도로 파산한 기업은 236으로 전체부도율은 γ=0.057이다. 예제 자료의 구성은 다음과 같다.

표 4.1 실증예제 자료구성

변수이름	변수설명	형태
Y	신용상태	범주형(0,1)
X_a	신용점수	연속형
X_b	차입금 의존도	연속형
X_c	금융비용비율 (금융비용/총부채)	연속형
X_d	자산대비 부채비율 (부채총계/자산총계)	연속형

4.1. 미결정구간이 존재하지 않는 경우

절단점을 X_s변수의 값 35으로 설정하고 이 평점이하의 기업에 대하여는 부도로 예측할 때, 실제와 예측한 신용상태에 대한 혼동행렬 (confusion matrix)는 표 4.2와 같다. 이에 대응하는 전체 오분류율은 15.48%이다.

표 4.2 전체 자료의 혼동행렬

전체		실제	
		부도	정상
결정	부도	204	606
	정상	34	3290

4.2. 미결정구간 설정한 경우1

미결정 구간을 $35 - 10.6 < X_s \leq 35 + 10.6$인 경우로 설정하자. 이 경우에는 미결정자가 전체 자료의 30%인 크기 $n_1 = 1246$이 되고, $X_1 \equiv 10.6 - |X_s - 35|$으로 표현하여 SAS의 PROC QLIM을 사용하여 얻은 이변량 프로빗모형에서의 선택모형은 다음과 같다.

$$\widehat{A}_i = \begin{cases} 1, & \boldsymbol{X}_{1i}\hat{\beta}_1 \geq 0 \\ 0, & \text{그 외}, \end{cases}$$

여기서 $\hat{\beta}_1 = (b_0, b_1)'$이며 b_0와 b_1은 각각 0과 1에 수렴한다. 그리고 표본크기 $n - n_1$의 결정된 차주들에 대하여 특성변수로 $\boldsymbol{X}_2 = (X_a, X_b, X_c)$로 설정할 때의 결정자에 대한 이변량 프로빗모형의 결과모형을 다음과 같이 얻는다. 임의의 결정자 $i = 1, \ldots, 2888$에 대하여,

$$\widehat{Y}_i = \begin{cases} 1, & \Phi\left(\boldsymbol{X}_{2i}\hat{\beta}_2\right) \geq 0.5 \\ 0, & \text{그 외}, \end{cases}$$

여기서 $\boldsymbol{X}_{2i}\hat{\beta}_2 = 6.282_{(0.366)} - 0.008_{(0.003)}X_{ai} - 0.049_{(0.004)}X_{bi} - 0.087_{(0.020)}X_{ci}$이고 회귀계수의 아래 괄호안은 표준오차를 나타낸다. 결정자에 대한 결과모형의 추정식을 사용하여 미결정자를 추론한다. 미결정자 차주 $j = 1, \ldots, 1246$에 대하여,

$$\widehat{Y}_j = \begin{cases} 1, & \widehat{P}\left(Y_j = 1 | \boldsymbol{X}_{2j}, A_j = 1\right) = \Phi\left(\boldsymbol{X}_{2j}\hat{\beta}_2\right) \geq 0.5, \\ 0, & \widehat{P}\left(Y_j = 1 | \boldsymbol{X}_{2j}, A_j = 1\right) = \Phi\left(\boldsymbol{X}_{2j}\hat{\beta}_2\right) < 0.5. \end{cases}$$

이 추론방법으로 미결정자를 추정한 후 미결정자에 대한 실제 Y와 추정된 Y사이의 혼동행렬과 전체 자료에 대한 혼동행렬은 표 4.3에 나타내었다. 미결정자에 대한 오분류율은 8.10%이고, 전체에 대한 오분류율은 7.42%이다.

표 4.3 MNAR 가정의 미결정자 추론결과와 추론 후 전체의 혼동행렬

미결정		실제		전체		실제	
		부도	정상			부도	정상
결정	부도	2	3	결정	부도	131	200
	정상	98	1144		정상	107	3696

$\boldsymbol{X}_2$에 추가된 정보인 $\boldsymbol{X}^+ \equiv X_d$를 포함한 새로운 특성변수 $\boldsymbol{X}_3 \equiv (\boldsymbol{X}_2, \boldsymbol{X}^+) = (X_a, X_b, X_c, X_d)$를 사용하여 결정자가 부도일 확률을 다음과 같이 추정한다. 임의의 결정자 $i = 1, \ldots, 2888$에 대하여,

$$\hat{P}(Y_i = 1 | \boldsymbol{X}_{3i}, A_i = 0) = \Phi(\boldsymbol{X}_{3i}\hat{\beta}_3)$$
$$= \Phi(2.447_{(0.747)} + 0.036_{(0.007)}X_{ai} - 0.005_{(0.004)}X_{bi} - 0.039_{(0.022)}X_{ci} - 0.022_{(0.006)}X_{di}).$$

미결정자의 우 · 불량 여부는 추정한 부도 확률이 $\Phi\left(\boldsymbol{X}_{3j}\hat{\beta}_3\right) \geq 0.5$일 때로 추정하여, 미결정자와 전체자료에 대한 혼동행렬을 표 4.4에 나타내었다. 미결정자 오분류율은 7.94%, 전체 오분류율은 7.37%이다. 특성변수 $\boldsymbol{X}_2 = (X_a, X_b, X_c)$를 사용하는 경우보다 $\boldsymbol{X}_3 = (X_a, X_b, X_c, X_d)$를 사용하였을 때, 미결정자의 오분류율이 8.10%에서 7.94%로 오분류율이 줄어들었음을 확인할 수 있다.

표 4.4 MNAR 가정의 미결정자 추론결과와 추론 후 전체의 혼동행렬

미결정		실제		전체		실제	
		부도	정상			부도	정상
결정	부도	3	2	결정	부도	132	199
	정상	97	1145		정상	106	3697

4.3. 미결정구간 설정한 경우2

미결정 구간을 $35 - 14.3 < X_s \leq 35 + 14.3$인 경우로 확대 설정하자. 그러면 미결정자를 전체자료의 40%인 수준으로 선정하고 $n_1 = 1654$이다. 특성변수로 $\boldsymbol{X}_1 \equiv 14.3 - |X_s - 35|$, $\boldsymbol{X}_2 = (X_a, X_b, X_c)$를 설정하고 이변량 프로빗모형에서의 선택모형과 결과모형을 앞에서와 같이 추정한다. 여기에서도 $\hat{\beta}_1 = (b_0, b_1)$의 b_0 와 b_1은 각각 0과 1에 수렴하고, 결정자에 대한 결과모형의 추정식은 다음과 같다. $i = 1, \ldots, 2480$에 대하여 $\boldsymbol{X}_{2i}\hat{\beta}_2 = 7.007_{(0.489)} - 0.009_{(0.004)}X_{ai} - 0.058_{(0.005)}X_{bi} - 0.063_{(0.026)}X_{ci}$이다. 결정자에 대한 결과모형의 추정식을 이용하여 미결정자에 대하여 추정한 후 미결정자에 대한 실제 Y와 추정된 Y사이 그리고 전체 자료에 대한 혼동행렬은 표 4.5에 나타내었다. 미결정자에 대한 오분류율은 8.04%이고, 전체에 대한 오분류율은 6.12%이다. 미결정구간이 30%에서 40%로 확대될 때, 오분류율이 8.10%에서 8.04%로 감소했음을 알 수 있다.

표 4.5 MNAR 가정의 미결정자 추론결과와 추론 후 전체의 혼동행렬

미결정		실제		전체		실제	
		부도	정상			부도	정상
결정	부도	4	7	결정	부도	108	119
	정상	126	1157		정상	130	3777

다음으로는 $\boldsymbol{X}_3 = (X_a, X_b, X_c, X_d)$를 특성변수를 사용하여 결정자의 부도일 확률을 다음과 같이 추정한다. 결정자 $i = 1, \ldots, 2480$에 대하여,

$$\begin{aligned}&\widehat{P}\left(Y_i = 1 | \boldsymbol{X}_{3i}, A_i = 0\right) = \Phi\left(\boldsymbol{X}_{3i}\hat{\beta}_3\right)\\ &= \Phi(3.574_{(0.954)} + 0.030_{(0.004)}X_{ai} - 0.006_{(0.004)}X_{bi} - 0.034_{(0.007)}X_{ci} - 0.016_{(0.028)}X_{di}).\end{aligned}$$

미결정자에 대하여는 확률이 $\Phi\left(\boldsymbol{X}_{3j}\hat{\beta}_3\right) \geq 0.5$일 때 부도로 판단하여, 미결정자와 전체자료에 대한 혼동행렬을 표 4.6에 나타내었다. 미결정자 오분류율은 7.55%, 전체 오분류율은 6.79%이다. 특성변수 $\boldsymbol{X}_2$를 사용한 표 4.5와 $\boldsymbol{X}_3$를 사용한 표 4.6을 비교했을 때, 미결정자 오분류율이 8.04%에서 7.55%로 감소하였음을 파악할 수 있다. 미결정 비율이 30%이며 특성변수 $\boldsymbol{X}_2$를 사용한 표 4.3의 미결정자 오분류율 8.10%에 비해 미결정 비율이 40%이며 특성변수 $\boldsymbol{X}_3$를 사용한 표 4.6의 미결정자 오분류율 7.55%로 감소했음을 알 수 있다. 미결정구간을 확대하면 정상 집단과 부도 집단의 정보를 더욱 정확하게 반영할 수 있다는 측면에서 미결정자 오분류율과 전체 오분류율이 줄어드는 것을 확인할 수 있다.

표 4.6 MNAR 가정의 미결정자 추론결과와 추론 후 전체의 혼동행렬

미결정		실제		전체		실제	
		부도	정상			부도	정상
결정	부도	8	3	결정	부도	112	115
	정상	122	1521		정상	126	3781

4.4. 미결정구간 설정한 경우3

절단점을 X_s=30으로 변경하고 미결정 구간을 $30-17.4 < X_s \leq 30+17.4$인 경우로 설정하여 살펴본다. 미결정 차주수는 전체자료의 40%인 $n_1 = 1655$이 된다. 특성변수로 $\boldsymbol{X}_1 \equiv 17.4 - |X_s - 30|$, $\boldsymbol{X}_2 = (X_a, X_b, X_c)$를 설정하고 이변량 프로빗모형에서의 선택모형과 결과모형을 4.2와 4.3절에서와 같이 추정하면, $\hat{\beta_1} = (b_0, b_1)$의 b_0 와 b_1은 각각 0과 1에 수렴하고, 결정자에 대한 결과모형의 추정식은 다음과 같다. $i = 1, \ldots, 2479$에 대하여 $\boldsymbol{X}_{2i}\hat{\beta}_2 = -0.295_{(0.501)} + 0.063_{(0.006)}X_{ai} - 0.014_{(0.005)}X_{bi} - 0.080_{(0.035)}X_{ci}$이다. 결정자에 대한 결과모형의 추정식을 바탕으로 미결정자 $j = 1, \ldots, n_1$에 대하여 추정한 후, 미결정자에 대한 실제 Y와 추정된 Y사이 그리고 전체자료에 대한 혼동행렬은 표 4.7에 나타내었다. 미결정자에 대한 오분류율은 12.44%이고, 전체에 대한 오분류율은 5.44%이다. 미결정자 비율이 40%로 동일할 경우, 표 4.5의 절단점이 35이며 $\boldsymbol{X}_2$를 사용한 오분류율 8.10%에 비해 표 4.7의 절단점이 30이며 $\boldsymbol{X}_2$를 사용한 오분류율이 12.44%로 증가하였다.

표 4.7 MNAR 가정의 미결정자 추론결과와 추론 후 전체의 혼동행렬

미결정		실제		전체		실제	
		부도	정상			부도	정상
결정	부도	46	71	결정	부도	96	83
	정상	135	1403		정상	142	3813

추가된 $\boldsymbol{X}_3 = (X_a, X_b, X_c, X_d)$를 특성변수로 사용하여 결정자 모형을 다음과 같이 추론한다. 결정자 $i = 1, \ldots, 2479$에 대하여,

$$\widehat{P}\left(Y_i = 1 | \boldsymbol{X}_{3i}, A_i = 0\right) = \Phi\left(\boldsymbol{X}_{3i}\hat{\beta}_3\right)$$
$$= \Phi(0.672_{(1.051)} + 0.054_{(0.010)}X_{ai} - 0.009_{(0.006)}X_{bi} - 0.023_{(0.038)}X_{ci} - 0.098_{(0.009)}X_{di}).$$

추정된 회귀계수 $\hat{\beta}_3$를 사용하여 $\Phi\left(\boldsymbol{X}_{3j}\hat{\beta}_3\right) \geq 0.5$이면 미결정자가 부도로 추정하고, 미결정자와 전체자료에 대한 혼동행렬을 표 4.8에 나타내었다. 미결정자 오분류율은 11.72%, 전체 오분류율은 5.15%이다. 특성변수 $\boldsymbol{X}_2$를 사용한 표 4.7의 오분류율이 12.44%이고, $\boldsymbol{X}_3$를 사용한 표 4.8의 오분류율이 11.72%로 감소하였음을 파악할 수 있다. 그러나, 절단점이 35인 경우의 표 4.6의 오분류율 7.55%에 비해서 11.72%는 조금 높은 수치이다.

표 4.8 MNAR 가정의 미결정자 추론결과와 추론 후 전체의 혼동행렬

미결정		실제		전체		실제	
		부도	정상			부도	정상
결정	부도	48	61	결정	부도	96	73
	정상	133	1413		정상	140	3823

5. 결론

결측자료 문제로 간주한 미결정자 추론은 심사가 이루어지지 않고 판단이 보류되어 발생하는 미결정자는 MAR 가정을 하고, 신용평가를 판단하기 어려운 평점 때문에 평가를 유보하고 특별한 전문가에게 재심사를 의뢰하기 위하여 결정이 보류되어 발생하는 미결정자는 MNAR 가정을 한다. 미결정자 그룹의 승인·거절 결정이 결정자 그룹의 특성변수의 영향력이 다르거나 특성변수 이외에 추가적으로 영향을 주는 요소가 존재하고, 동일한 속성을 갖더라도 미결정자의 우량 확률과 미결정을 제외한 집단의 우량 확률은 다르게 계산되므로 MNAR 가정이 필요하다. 본 연구에서는 결측자료 유형 중 MNAR 가정 하에서 이변량 프로빗모형을 이용한 미결정자 추론방법을 제안하였다. 결정자 집단과 미결정자 집단의 선택과정이 결정자 집단의 신용평가모형을 추정하는 과정에 영향을 주고 특성변수를 고려하여 미결정자를 예측하는데, 선택모형과 결과모형 그리고 오차항의 상관계수를 고려하는 이변량 프로빗모형을 이용하는데 큰 의의가 있다.

본 논문에서 두 가지 미결정자 추론방법을 제시하였다. 차주의 특성을 나타내는 특성변수를 사용하여 미결정 차주의 부도여부를 추론하는 방법과 미결정 차주의 신용평가가 어려우므로 더욱 정확한 정보를 추가적으로 수집한 후 새로운 특성변수를 사용하여 미결정 차주의 부도여부를 추론하는 방법을 제안하였다.

자산규모가 4,500억 이상인 한국 외감기업인 총 4,134 기업에 대한 자료에서 미결정 구간을 임의로 설정하고 제안한 방법을 사용하여, 특성변수의 변화와 미결정 구간의 변화, 그리고 절단점의 변화에 따라 미결정자와 전체 오분류율을 비교한 결과가 표 5.1과 같다.

표 5.1 오분류율 비교

미결정 비율 (절단점)	미결정자 30% (절단점 : 35)		미결정자 40% (절단점 : 35)		미결정자 40% (절단점 : 30)	
특성변수	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$
미결정자	8.10%	7.94%	8.04%	7.55%	12.44%	11.72%
전체	7.42%	7.37%	6.12%	6.79%	5.44%	5.15%

표 5.1에서 $\boldsymbol{X}_2$를 고려하였을 때보다 $\boldsymbol{X}_3$를 고려하였을 때 미결정 비율이 30%이며 절단점이 35인 경우 미결정자에 대한 오분류율은 8.10%에서 7.94%로, 미결정 비율이 40%이면서 절단점이 35인경우 미결정자에 대한 오분류율은 8.04%에서 7.55%로, 미결정 비율이 40%이면서 절단점이 30인 경우 12.44%에서 11.72%로 미결정자 오분류율이 감소함을 알 수 있다. 더욱 정확한 신용정보를 모형에 추가하여 사용하였기 때문에 오분류율을 줄일 수 있다고 판단된다. 그리고 절단점이 35로 동일할 경우 미결정자 비율이 30%에서 40%로 증가하면 특성변수 $\boldsymbol{X}_2$일 때의 오분류율이 8.10%에서 8.04%로, 특성변수 $\boldsymbol{X}_3$일 때는 7.94%에서 7.55%로 줄어드는 것을 알 수 있다. 그러므로 추가적인 특성변수 $\boldsymbol{X}_3$를 고려하여 미결정자를 추정했을 경우의 오분류율이 특성변수 $\boldsymbol{X}_2$를 고려하였을 때보다 감소하는 것을 확인할 수 있었다. 또한 미결정구간을 확대하면, 정상 집단과 부도 집단의 정보를 더욱 정확하게 반영할 수 있기 때문에 미결정자와 전체 오분류율이 더욱 큰 감소효과가 발생한다고 파악된다.

절단점을 오분류율이 더욱 커지는 위치로 설정했을 경우에 미결정자의 오분류율은 증가하지만, 특성변수 $\boldsymbol{X}_2$를 사용할 때보다 $\boldsymbol{X}_3$를 고려하여 미결정자를 추정했을 경우의 미결정자의 오분류율이 12.44%에서 11.72%로 감소하는 것을 발견하였다. 절단점 선정과 오분류율의 변화에 대하여는 향후 연구과제로 남겨놓기로 한다.

참고문헌

홍종선, 정민섭 (2011). 신용평가에서 로지스틱회귀를 이용한 미결정자 추론. <한국데이터정보과학회지>, **22**, 149-157.

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

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 incricket. *Journal of Quantitative Analysis in Sports*, **6**, 1-13.

Feelders, A. J. (2000a). *An overview of model based reject inference for credit scoring*, Utrecht University, Institute for Information and Computing Sciences.

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

Greene, W. H. (1996). *Marginal effects in the bivariate probit model*, NYU Working Paper No. EC-96-11.

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

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, **47**, 153-161.

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.

Meng, C. and James R. V. (2002). A statistical model of bilateral cooperation. *Political Analysis*, **10**, 101-112.

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.

Poirier, D. J. (1980). Partial observability in bivariate probit models. *Journal of Econometrics*, **12**, 210-217.

Sartori, A. (2003) An estimator for some binary-outcome selection models without exclusion restrictions. *Ploitical Analysis*, **11**, 111-138.

Journal of the Korean Data & Information Science Society
2011, **22**(6), 1017–1028

한국데이터정보과학회지

Undecided inference using bivariate probit models

Chong Sun Hong[1] · Mi Hyang Jung[2]

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

Received 14 September 2011, revised 18 October 2011, accepted 23 October 2011

Abstract

When it is not easy to decide the credit scoring for some loan applicants, credit evaluation is postponded and reserve to ask a specialist for further evaluation of undecided applicants. This undecided inference is one of problems that happen to most statistical models including the biostatistics and sportal statistics as well as credit evaluation area. In this work, the undecided inference is regarded as a missing data mechanism under the assumption of MNAR, and use the bivariate probit model which is one of sample selection models. Two undecided inference methods are proposed: one is to make use of characteristic variables to represent the state for decided applicants, and the other is that more accurate and additional informations are collected and apply these new variables. With an illustrated example, misclassification error rates for undecided and overall applicants are obtainded and compared according to various characteristic variables, undecided intervals, and thresholds. It is found that misclassification error rates could be reduced when the undecided interval is increased and more accurate information is put to model, since more accurate situation of decided applications are reflected in the bivariate probit model.

Keywords: Credit evalution, missing data, probit model, reject inference, sample selection.

[1] Corresponding author: Professor, Department 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 and Graduate student, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea.