Journal of the Korean Data Analysis Society (February 2017)
Vol. 19, No. 1 (B), pp. 129-139

결측이 있는 신용평가 자료의 희소사건 로지스틱 회귀분석*

이보은[1], 주용성[2], 전형준[3]

요 약

기업이나 개인의 파산여부에 대한 예측은 은행의 대출에 관한 의사결정시 주요 지표로 활용되어왔다. 본 논문에서는 파산여부 자료의 통계 분석과정에서 주의를 필요로 하는 두 가지 특징을 가진다. 첫째, 자료의 종속변수인 파산여부는 희소사건이다. 희소사건의 발생확률을 예측하기 위하여 로지스틱 회귀모형에 일반적인 최대우도 추정법을 적용하는 경우 회귀계수 추정치에 편향이 발생하여 사건의 예측확률이 과소추정 되는 경향을 가진다. 대부자의 입장에서는 파산가능성을 과소추정하여 상환불이행 등을 통한 손실을 입는 가능성이 높아지게 된다. 둘째, 파산여부 자료에는 결측이 존재할 가능성이 높다. 결측치가 없는 완전자료를 사용하는 경우 자료의 손실로 인한 모형의 정확도가 감소하고 모형 추정과정에서 편향이 발생할 수 있다. 본 논문에서는 파산여부 예측모형에서 발생할 것으로 예상되는 편의들의 보정방법과 모의실험을 통한 보정방법의 효용성을 검증하였다.

주요용어 : 신용평가, 희소사건, 로지스틱 회귀분석, 결측치 대체, 편의 보정.

1. 서론

지난 IMF 외환위기를 단시간 내에 극복하기 위하여 정부에서는 무리한 경기부양책으로 카드에 대하여 지나친 규제완화를 단행하였다. 결과적으로 신용카드대란이 발생하였고, 금융위기 직후인 2009년 1/4분기에도 8.3%까지 올라갔던 청년(20~29세)실업률은 2016년 1/4분기 기준 11.2%까지 올라갔다(KOSIS, 2017). 최근에도 가계의 건전성을 위협하고 있는 가계부채 폭등 문제가 지금 우리 경제의 시한폭탄이라는 지적이 있다. 정부는 2015년 기준으로 가계부채가 1,224조(=가구당 평균 부채액 6,256만원×1,956만 가구)(KOSIS, 2017)라고 발표하였다. 그 규모뿐만 아니라 증가 속도도 너무 빠르고, 가처분 소득 대비 가계부채 비율이 2015년도에 170%(OECD, 2017)를 육박하여 매우 위험한 수준이다. 미국이 지난 금융위기를 2008년도에 겪을 때, 143%(OECD, 2017)가 채 안 되었던 것과 비교하면 상당히 심각한 수준이다. 특히 박근혜정부의 부동산 활성화 대책을 중심으로 한 소위 초이노믹스는 가계부채 문제를 악화시켰다는 게 대부분 전문가들의 의견이다. 현재의 가계부채를 이대로 방치하면 과거 노무현 정부 때의 신용카드대란사태를 훨씬 상회하는 신용불량자가 양산될 수 있으므로 대책마련과 예측을 위한 정확한 통계모형의 개발이 절실하다.

파산여부 예측모형은 Hand(1997)가 기계학습(machine learning)기법을 통해 개발한 것을 시작으로 많은 연구자들을 통해 지금까지 발전해오고 있다. 기업과 개인의 파산여부 가능성은 은행이 개인

*이 논문은 2015-2016년도 창원대학교 자율연구과제 연구비지원으로 수행된 연구결과임.
*이 논문은 제1저자 이보은의 석사학위논문의 계속연구로 작성한 것입니다.
[1]04620 서울특별시 중구 필동로1길 30, 동국대학교 통계학과 석사과정. E-mail : leebe9112@dongguk.edu
[2]04620 서울특별시 중구 필동로1길 30, 동국대학교 통계학과 교수. E-mail : yongsungjoo@dongguk.edu
[3](교신저자) 51140 경상남도 창원시 의창구 창원대학로 20, 창원대학교 세무학과 교수.
E-mail : jeon266@changwon.ac.kr
[접수 2017년 1월 20일; 수정 2017년 2월 7일, 2017년 2월 17일; 게재확정 2017년 2월 20일]

의 대출실행이나 카드개설 등의 의사 결정시 활용하기 위한 참고지표로 유용하게 사용되어 왔다. 개인신용등급은 신용평가 회사마다 비중의 차이는 있지만 해당 개인의 채무 적시 상환 여부, 현재 채무 수준, 신용거래 기간, 신용활용의 정도를 평가요소로 하여 결정한다. 개인신용평가회사인 나이스신용평가정보와 코리아크레딧뷰로는 개인의 과거 부채상환기록, 현재의 부채수준, 연체정보, 제2금융권(저축은행, 캐피탈) 과다 이용 여부 및 단기간(3~6개월) 집중 대출 여부 등을 종합하여 개인의 신용등급을 평가한다.

기업이나 개인의 파산여부 자료는 일반적인 통계기법으로 다루기 힘든 두 가지 특징을 가지고 있다. 첫째, 파산은 쉽게 일어나지 않는 희소사건이다. 파산여부 예측모형 개발에 로지스틱 회귀분석 모형이 널리 사용되는데(Kwak et al., 2015; Lee, Song, 2007; Park, Bang, 2015) 이는 모형의 예측력과 설명력이 높기 때문이다(Kim, Kang, Kim, 2012). 하지만 파산여부 예측모형의 종속변수가 희소사건인 경우 일반적인 로지스틱 회귀분석을 사용하면 회귀계수 추정량에 편향(bias)이 발생하며 사건발생확률을 과소추정 하는 문제점이 있다(King, Zeng, 2001). 따라서 일반적인 모형 적합결과를 그대로 사용하게 되면 대출자의 파산확률을 과소 예측하거나 혹은 신용을 과대 예측하게 된다. 희소사건에 기인한 문제점을 해결하는 대표적인 방법으로 King, Zeng(2001)이 제시한 방법과 Firth (1993)가 제시한 방법이 있다.

둘째, 파산여부 자료에서 특정 개인이나 기업의 내부정보 접근성 문제, 자료 수집자(개인이나 회사 혹은 기관)들의 상이한 수집방식 등의 이유로 결측이 흔히 발생할 수 있다. 결측이 있는 자료를 처리하는 가장 단순한 방법으로 결측이 없는 기업이나 개인의 완전한 자료만을 이용하여 분석하는 완전제거법(complete case analysis)을 고려할 수 있지만 이 방법은 자료에 큰 손실을 가져올 수 있고 결측메커니즘에 따라 편향이 발생할 수 있다. 이러한 단점을 극복하기 위하여 적절한 하나의 값으로 결측치를 대체하는 단일대체(single imputation)와 여러 개의 대체값을 생성하여 분석하는 다중대체(multiple imputation)방법이 사용되고 있다(Rubin, 1987; Little, Rubin, 2002). 결측패턴이 일반적(general)이고 결측메커니즘이 임의결측(missing at random) 가정일 경우 조건부분포를 이용하여 결측값을 대체하는 완전 조건부(fully conditional specification)알고리즘이 다중대체방법으로 널리 사용되고 있다(Van Buuren, 2007; Lee, Carlin, 2010). 본 연구에서는 결측치를 대체한 후 희소사건에 따른 편의를 보정한 로지스틱 회귀분석 방법의 유효성검증을 위한 모의실험을 실시하였다.

본 논문의 2절에서 결측메커니즘과 결측패턴을 소개하고 3절에서는 결측치 대체방법을 소개하였다. 4절에서는 종속변수가 희소사건인 로지스틱 모형 추정방법에 대해 설명하였다. 5절에서는 희소사건을 포함한 결측자료의 결측비율와 희소사건발생률을 조절하는 모의실험을 통해 편의 보정방법들을 비교하였다. 모의실험에서는 임의의 자료를 생성하는 대신, 프랑스 기업자료(Yu et al., 2009; Aalto University, 2017)를 복원추출하여 실제 신용평가 자료의 특성을 반영하는 모의자료를 생성하였다.

2. 결측치의 발생방식 및 결측패턴

결측값은 실험이나 조사에서 다양한 이유로 관측되어야 할 값을 얻지 못한 경우를 말하며 파산여부예측을 위한 신용평가 자료에서 종종 접할 수 있다. 결측자료 메커니즘은 결측여부와 변수간의 관계를 나타내는 것이며, 관련된 정도에 따라 완전임의결측, 임의결측, 비임의결측으로 분류할 수 있다(Little, Rubin, 2002). 완전임의결측(missing at completely random)이란 결측이 발생할 확률은 자료와는 관련 없이 독립인 경우를 말한다. 이 경우 완전제거법을 사용하더라도 추정치에 편향이 발생하지 않아 유효한 결론을 얻을 수 있다. 하지만 완전임의결측은 현실적으로 희박하게 발생하

므로 대부분의 대체방법들은 임의결측(missing at random)가정을 기본으로 하고 있다. 임의결측은 결측이 발생할 확률이 관측된 자료에 영향을 받지만 결측된 자료와는 무관하게 발생하는 경우를 말한다. 따라서 완전자료를 가진 어떤 변수가 결측이 발생한 변수와 관련이 있는지 파악하는 것이 결측 자료를 분석하는데 효과적이라 할 수 있다. 비임의결측(not missing at random)은 결측이 발생할 확률이 결측된 자료와 관련이 있는 경우를 말하며, 관측된 자료를 통해 결측 여부를 예측할 수 없다는 것을 의미한다. 결측 메커니즘이 완전 임의결측과 임의결측일 경우 본 논문에서 사용되는 이론으로 결측 자료에 적용할 대체 방법이 활용될 수 있으나, 비임의결측일 경우 앞의 두 메커니즘보다 복잡한 대체과정이 요구되고 추정치의 편향이 발생할 수 있다.

3. 결측대체

결측을 다루는 방법에는 크게 결측치를 제거하여 결측치가 없는 완전한 자료만을 사용하는 완전제거법, 결측이 존재하는 자료를 그대로 사용하는 방법과 결측이 대체된 자료를 사용하는 방법이 있다(Little, Rubin, 2002). 첫째, 임의결측 가정하에서의 완전제거법은 관측된 자료와 결측된 자료의 분포가 다를 수 있으므로 추정의 편향을 발생시킨다. 둘째, 결측이 존재하는 자료를 그대로 사용하는 방법에는 부트스트랩이나 잭나이프(Meng, Rubin, 1991)와 같은 재추출(resampling) 방법이 있다. 셋째, 추정량의 편향을 개선하기 위한 대체방법에는 단일대체법(single imputation)과 다중대체법(multiple imputation)이 있다. 단일대체방법 중 평균대체나 중앙값 대체, 결정론적(deterministic) 회귀대체방법(Gelman, 2006)은 미지의 결측값을 예측하는데 불확실성(uncertainty)을 반영하지 못하여 자료의 분산을 과소추정하는 문제점이 있다(Rubin, 1987; Pastor, 2003). EM대체방법과 확률론적(random) 회귀대체 방법은 자료 분산의 과소추정문제를 해결한 반면, 표준오차를 구하기 위해서는 추가적으로 복잡한 수리적 계산 혹은 수치해석을 필요로 한다(Jamshidian, Jennrich, 2000; Gelman, Hill, 2006). 결과적으로 변수간의 연관성이 복잡한 경우 정확한 표준오차를 구하기 어렵게 된다. 단일대체법과는 달리 다중대체법은 결측치가 대체된 자료를 임의로 반복 생성함으로써 표준오차를 쉽게 구할 수 있도록 한다.

다중대체법은 자료의 결합확률분포를 다변량 정규분포로 가정하는 방법과 좀 더 일반적인 경우에 적용 가능한 완전 조건부(FCS: fully conditional specification) 알고리즘으로 나눠질 수 있다. FCS 방법은 시간이 오래걸리는 단점이 있으나, 조건부 분포를 이용하여 대체값을 구하므로 자료의 결합확률분포를 통해 구하는 방법보다 분석이 간편하다. 또한 안정적인(robust) 결과를 낼 수 있으므로(Van Buuren et al., 2006) 본 논문에서는 FCS방법을 이용하여 분석하였다. FCS 알고리즘은 공변량에 이산형과 연속형이 함께 존재하는 경우에 사용이 가능하다(Van Buuren, 2007; Lee, Carlin, 2010).

다중대체방법은 반복적인 임의추출을 이용한 알고리즘이기 때문에 분석시간이 오래 걸리는 대신, 표준오차의 추정이 쉽다. 반면, 단일대체법은 표준편차를 추정하기 어려운 대신, 모수의 평균 혹은 회귀함수의 추정을 위한 분석시간이 훨씬 짧기 때문에 대용량의 분석에도 활용가능한 장점이 있다.

FCS 알고리즘은 각 변수의 조건부 분포를 연속적으로 적용하는 방법으로 깁스샘플링(Casella, George, 1992)과 유사한 형태를 가진다. FCS 알고리즘은 두 단계로 구성된다. 첫 번째 단계에서 결측이 존재하는 변수에서 관측된 값을 임의로 추출하여 초기 대체값으로 설정한다. 두 번째 단계는 결측이 존재하는 각 변수를 종속변수로 사용하는 일반화선형모형을 연속적으로 사용한다. 종속변수의 값이 원자료에서 결측인 경우 이를 예측값으로 대체하고 대체값들이 수렴에 가까워질 때까지

이 단계를 반복한다. 즉, 두 번째 단계는 모형의 예측값을 구하듯이 결측값을 채우고 앞서 채워진 변수가 다음 채워지는 변수의 독립변수로 활용되는 형태이다(Van Buuren et al., 2006; Buuren, Groothuis Oudshoorn, 2011).

4. 희소 사건 로지스틱 모형 추정

사건이 발생하는 경우가 발생하지 않는 경우에 비해서 상대적으로 매우 적은 경우 이를 희소 사건이라 한다. 반응변수의 사건발생률이 매우 낮은 희소사건의 경우 기존의 로지스틱 회귀분석의 회귀계수 추정치에 편향이 발생하고 사건발생확률을 과소 추정 한다고 알려져 있다(King, Zeng, 2001). 게다가, Figure 1에서 볼 수 있듯이 특정 x값들에 의해 사건이 준완전분리(quasi-complete separation) 혹은 완전분리(complete separation)되는 상황이 쉽게 발생할 수 있다. 일반적인 최대우도 모수 추정치는 발산에 가까워지거나 발산하여 올바른 추정값을 얻을 수 없게 된다. 회귀계수 편의 보정을 위한 방법들이 Firth(1993)와 King, Zeng(2001)에서 소개되었다.

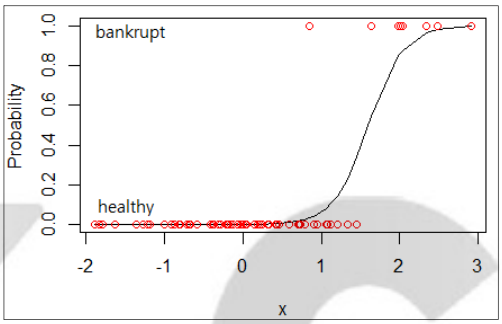

Figure 1. Rare event logistic curve(Lee, 2017)

4.1. Firth 방법

로지스틱 회귀모형은 반응변수 Y_i, $i=1,\cdots,n$가 서로 독립적으로 이항분포 $B(m_i,p_i)$를 따를 때 사용하는 대표적인 모형이다. 최대우도 추정량(maximum likelihood estimator)을 구하기 위한 스코어 함수(score function)는 식 (4.1)과 같다.

$$\ell(\underline{\beta}) = \sum_{i=1}^{n}\left\{y_i(\underline{x_i}^T\underline{\beta}) - m_i\log\left(1+\exp(\underline{x_i}^T\underline{\beta})\right)\right\}$$
$$U(\beta_i) = \frac{\partial\ell(\underline{\beta})}{\partial\beta_j} = \sum_{i=1}^{n}\left\{x_{ij}\left(y_i - m_i\frac{\exp(\underline{x_i}\underline{\beta})}{1+\exp(\underline{x_i}\underline{\beta})}\right)\right\},\quad j=0,1,\cdots,r \tag{4.1}$$

여기에서 m_i는 각 변수 i의 관측수, β_0은 절편, β_r은 회귀계수, x_{ir}은 설명변수를 의미한다. $\ell(\underline{\beta})$는 로그우도함수이고 $\underline{x_i}=(1\,x_{i1}\,\cdots\,x_{ir})$, $\underline{\beta}=(\beta_0,\beta_1,\cdots,\beta_r)$이다. 로지스틱 모형의 스코어함수 식 (4.1)은 계수에 대한 비선형 함수이므로 식 (4.2)와 같이 뉴튼-랩슨(Newton-Raphson) 알고리즘을 이용하여 구할 수 있다. 식 (4.2)의 $I(\underline{\beta})$는 피셔의 정보행렬(Fisher's information matrix)로, $I(\underline{\beta}) = X^TWX$, $p_i = \exp(\underline{x_i}\underline{\beta})/(1+\exp(\underline{x_i}\underline{\beta}))$, $W = diag\{p_i(1-p_i)\}$이며 $I(\underline{\beta})$는 최대우도 추정량의 분산의 역함수와 일치한다. 일반적인 로지스틱 회귀모형의 추정방법에서는 식 (4.2)를 이용한다. 여기에서 $\underline{\beta}^{(s)}$는 s번째 뉴튼-랩슨 알고리즘의 반복을 의미한다(King, Zeng, 2001).

$$\underline{\beta}^{(s+1)} = \underline{\beta}^{(s)} + I^{-1}(\underline{\beta}^{(s)})\, U(\underline{\beta}^{(s)}) \tag{4.2}$$

반면, Firth(1993)은 희소사건 일반화 선형 모형에서 볼 수 있는 편향을 줄이기 위해 기존의 로그우도함수에 벌점을 부여하여 변형된 로그우도함수의 사용을 다음과 같이 제안하였다.

$$\log L(\underline{\beta})^{*} = \ell(\underline{\beta}) + \frac{1}{2} log\{\det(I(\underline{\beta}))\}$$

변형된 로그우도함수를 β_j에 대해 미분한 스코어함수는 식 (4.3)으로 표현된다. 여기서 h_i는 $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$의 i번째 대각원소이다. 변형된 스코어 함수 $U(\beta_j)^{*}$를 사용한 뉴튼-랩슨 알고리즘을 통하여 수정된 추정량을 구할 수 있다(Firth, 1993).

$$U(\beta_j)^{*} \equiv U(\beta_j) + \frac{1}{2} tr\left(I(\underline{\beta})^{-1} \frac{\partial \log I(\underline{\beta})}{\partial \beta_j} \right) = \sum \left\{ y_i - p_i + h_i \left(\frac{1}{2} - p_i \right) \right\} x_{ir} \tag{4.3}$$

4.2. King과 Zeng 방법

Firth(1993)가 제시한 방법과 달리 King, Zeng(2001)은 기존의 로지스틱 회귀분석을 통해 얻은 최대우도 추정량에 McCullagh, Nelder(1989)의 편의식을 빼서 편의보정한 추정량인 $\hat{\underline{\beta}}^{*}$를 얻는 방법을 다음과 같이 제시하였다. 여기에서 $bias(\hat{\underline{\beta}}) = (X^TWX)^{-1}X^TW\zeta$이고, $W\zeta$는 식 (4.3)의 $h_i(p_i - 1/2)$을 의미한다.

$$\hat{\underline{\beta}}^{*} = \hat{\underline{\beta}} - bias(\hat{\underline{\beta}})$$

King, Zeng(2001)이 제시하는 방법은 최대우도 추정치를 구한 후 그 값에서 추정치를 보정하므로 최대우도 추정치가 발산하는 경우(자료의 완전분리가 일어나는 경우)에는 유용하지 않다. 반면 Firth(1993)가 제시한 방법은 Jeffrey 사전분포(Jeffrey, 1946)를 이용하여 우도함수를 보정함으로써 완전분리(complete separation)가 발생하여도 추정치를 구할 수 있게 한다. 즉, King, Zeng(2001)의 방법은 편의를 수정(corrective)하는 반면 Firth(1993)의 방법은 편의를 예방(preventive)한다는 점에 차이가 있다(Heinze, Schemper, 2002).

5. 모의실험

결측이 흔히 존재하고 종속변수가 희소사건인 특징을 가지는 파산여부자료의 분석에서 결측치 대체와 로지스틱 회귀모형의 편의보정 효과를 보이도록 실제자료를 바탕으로 한 모의실험을 실시하였다. 모의실험에서 사용된 프랑스 기업들의 파산여부자료(Yu et al., 2009; Aalto University, 2017)는 공개된 데이터로, 500개 기업의 재무 프로파일을 나타내는 42개 변수로 이루어져 있으며 결측치는 존재하지 않는다. 종속변수인 기업의 파산여부는 기업의 신용이 안정적인 경우 'healthy'(250개)로, 기업이 파산한 경우인 'bankrupt'(250개)로 분류되어 있다. 연구목적으로 제공된 것이기 때문에 안정적인 기업의 수와 파산한 기업의 수가 같게 조정되어 있고 결측도 존재하지 않는다. 현실에서 파산은 희귀사건이고 조사 대상 기업의 정보는 결측을 포함하기 쉽다는 것을 반영하기 위하여 본 모의실험에서는 파산기업이 적게 임의추출 되도록 설정하였다. 파산여부 자료가 결측이 존재하고 파산이라는 희소사건 모의실험에 앞서 기업의 파산여부에 영향을 줄 가능성이 있는 변수들을 일반적인 로지스틱 회귀분석 변수선택법(후진제거법)을 통해 변수 5개를 선별(Table 1 참조)하였

다. Table 1의 EF3(총매출 대비 이자및세전이익), SE7(총자산 대비 총부채), LI2(총자산 대비 현금보유비율), RO6(총매출 대비 유동자산비율), API(자기자본의 변화)는 부채와 직접적으로 연관된 기업의 변제능력을 나타내는 변수들이다. 반면, 총매출 대비 부가가치세 비율이나 총매출 대비 노무비 비율과 같은 기업의 기타재무구조에 관한 변수들은 선택되지 않았다. 선별된 변수들과 파산여부 변수로만 자료를 재구성하여 전체자료로 사용하였다. 전체자료의 최대우도 추정방법을 적용하여 얻은 파산확률 예측식은 식 (5.1)과 같다.

$$\text{기업의 파산확률} = (1+e^{-A})^{-1}$$
$$A = \hat{\beta}_0 + \hat{\beta}_1 \times EF3 + \hat{\beta}_2 \times SE7 + \hat{\beta}_3 \times LI2 + \hat{\beta}_4 \times RO6 + \hat{\beta}_5 \times API \quad (5.1)$$

Table 1. Variable description

Variable	Description
EF3	EBIT / Total sales
SE7	Total debt / Total assets
LI2	Cash / Total assets
RO6	Current assets / Total sales
AP1	Change in equity position

5.1. 모의실험 설계

본 논문의 모의실험에서는 전체자료에 결측과 희소사건을 의도적으로 발생시켜 모의자료를 생성하였다. 모의자료로부터 추정된 회귀계수가 전체자료를 이용하여 추정된 회귀계수와 비슷한 값을 가질 수 있도록 추정방법을 개선할 수 있는지를 점검하였다. 모의실험의 자세한 진행방법은 다음과 같다.

현실에 부합하도록 전체자료에 변수마다 임의로 결측을 발생시켰고 종속변수인 기업의 파산여부를 희소사건에 해당하는 자료로 만들어주었다. 원자료의 500개 기업으로부터 파산 혹은 파산하지 않은 총 250개의 기업을 복원 추출하고 한 행에 한 개의 결측이 발생되도록 하였다. 파산발생률(파산된 회사의 비율)을 q, 결측비율(결측값을 가지는 기업의 비율)을 m이라 하자. 관심 있는 사건인 'bankrupt'가 현실에서 희소사건이기 때문에 파산여부가 'bankrupt'인 250개 기업 중 $250\times q$만큼 임의 추출하고, 파산 여부가 'healthy'인 기업 중에서 $250\times(1-q)$만큼 임의 추출하여 모의자료가 250개 기업을 포함하도록 자료를 구성한다. 파산기업의 구성비를 나타내는 q는 모의실험에서 1%, 3%, 5%, 10% 로 설정하여 희소의 정도가 심한 경우와 약한 경우가 비교되도록 하였다. 종속변수를 제외한 각 변수에 (자료수×m)/변수개수 만큼의 결측이 임의로 발생하게 한다. 위에서 만든 자료로 완전제거법을 적용하거나 FCS 알고리즘(Van Buuren, 2007; Lee, Carlin, 2010)을 이용하여 500번 다중대체법을 실시한다. 결측치 처리된 자료에서 파산여부 예측모형을 만든다. 이때 종속변수인 파산여부는 희소사건이므로 일반적인 로지스틱 회귀모형 추정법과 편의 보정방법들을 각각 적용하여 분석한다. 위의 모의실험을 30,000번 반복시행 하였다.

결측치 보정과 편의 보정 로지스틱 회귀분석의 효과를 측정하기 위하여 5개 설명변수만 포함하는 전체 프랑스 기업 자료(Yu et al., 2009; Aalto University, 2017)를 통해 구해진 회귀계수 추정치와 전체자료에 결측을 생성하여 분석한 회귀계수 추정치의 차이를 비교하였다. 희소사건을 포함하는 모의자료에서는 완전분리(complete separation)된 종속변수 때문에 추정치가 발산하게 되는 경우가 많이 발생한다. 모의자료들의 추정치 평균은 결과적으로 발산하여 방법에 따른 차이를 비교불

가능하게 한다. B는 모의실험 반복의 수, p는 회귀모형에서 사용되는 설명변수의 개수라고 하자. 일반적인 모의실험에서는 모의자료들의 추정치 평균과 목표하는 회귀계수값의 차이 $E(\widehat{\beta_{ij}})-\beta_{ij}^{*}$, $i=1,\cdots,B,\ j=0,\cdots,p-1$와 몬테카를로 분산 $\sum\left(E(\widehat{\beta_{ij}})-\beta_{ij}^{*}\right)^2/B$을 구하여 여러 가지 방법들을 비교하지만, 본 논문에서는 다음과 같이 중위수를 이용한 척도를 사용하였다. $\widehat{\beta}_{\cdot j}^{(sim)}$은 반복적으로 시행된 모의자료를 이용한 회귀모수 β_j의 추정치라고 하고, $\widehat{\beta}_j^{(whole)}$은 전체자료를 이용한 회귀계수 β_j의 추정치들의 벡터라고 하자. 또한, 중위수를 계산하는 함수를 med(·)라고 하자. 로지스틱 회귀계수 추정방법들을 비교하기 위하여 편향을 대신한 $D_j^{(med)}=med(\widehat{\beta}_{\cdot j}^{(sim)})-\widehat{\beta}_j^{(whole)}$를 사용하였으며, 몬테카를로 분산을 대신하여, $AD_{j}^{(med)}=med\left|\widehat{\beta}_{\cdot j}^{(sim)}-\widehat{\beta}_j^{(whole)}\right|$를 사용하였다.

5.2. 분석 결과

Table 2, Table 3의 'Missing rate'는 자료의 결측률을 나타내고 'Event rate'는 파산발생률을 나타

Table 2. Comparison of logistic regression estimation algorithm with $D_0^{(med)}$ and $AD_0^{(med)}$

Missing rate			10%		30%		50%	
Event rate			Complete case	Imputation	Complete case	Imputation	Complete case	Imputation
1%	MLE	$D_1^{(med)}$	-130.12	-115.59	-141.01	-130.90	-304.75	-153.80
		$AD_1^{(med)}$	129.60	129.06	131.52	130.38	183.39	159.23
	King	$D_1^{(med)}$	21.55	16.88	45.24	39.60	68.50	50.01
		$AD_1^{(med)}$	12.78	19.38	21.39	19.69	22.39	22.06
	Firth	$D_1^{(med)}$	-5.11	-4.02	-5.78	-4.30	-6.01	-4.45
		$AD_1^{(med)}$	0.57	0.40	0.80	0.79	0.79	0.99
3%	MLE	$D_1^{(med)}$	-125.59	-102.90	-135.37	-128.50	-277.86	-136.36
		$AD_1^{(med)}$	86.28	86.00	103.67	99.24	118.24	111.83
	King	$D_1^{(med)}$	16.88	14.28	41.03	31.21	57.77	46.46
		$AD_1^{(med)}$	17.39	18.00	25.99	23.05	36.15	29.24
	Firth	$D_1^{(med)}$	-2.99	-1.86	-3.12	-2.19	-3.53	-2.78
		$AD_1^{(med)}$	1.29	0.28	2.28	1.93	2.57	2.16
5%	MLE	$D_1^{(med)}$	-16.04	-15.62	-17.01	-15.69	-17.55	-16.16
		$AD_1^{(med)}$	16.20	16.13	43.29	40.18	118.62	80.46
	King	$D_1^{(med)}$	-5.42	-4.84	-5.68	-5.14	-6.03	-5.22
		$AD_1^{(med)}$	11.28	5.35	10.00	9.25	19.45	13.88
	Firth	$D_1^{(med)}$	-2.05	-1.09	-2.15	-2.01	-2.30	-2.11
		$AD_1^{(med)}$	1.90	1.74	2.38	2.00	1.60	1.46
10%	MLE	$D_1^{(med)}$	-7.59	-7.36	-7.90	-7.41	-8.69	-8.10
		$AD_1^{(med)}$	8.42	7.26	10.71	9.47	15.48	18.26
	King	$D_1^{(med)}$	-4.47	-4.09	-5.13	-4.12	-5.51	-4.52
		$AD_1^{(med)}$	9.39	8.11	10.38	10.23	10.47	10.37
	Firth	$D_1^{(med)}$	-1.12	-0.92	-1.53	-1.23	-2.01	-1.62
		$AD_1^{(med)}$	0.29	0.20	0.30	0.28	0.30	0.30

Table 3. Comparison of logistic regression estimation algorithm with $D_1^{(med)}$ and $AD_1^{(med)}$

Missing rate			10%		30%		50%	
Event rate			Complete case	Imputation	Complete case	Imputation	Complete case	Imputation
1%	MLE	$D_1^{(med)}$	-412.08	-405.84	-455.10	-436.91	-535.81	-492.72
		$AD_1^{(med)}$	394.08	240.30	311.21	255.13	315.01	289.91
	King	$D_1^{(med)}$	121.57	111.83	126.06	123.48	295.75	268.51
		$AD_1^{(med)}$	60.69	52.35	54.24	50.69	54.29	32.89
	Firth	$D_1^{(med)}$	-10.32	-9.21	-16.22	-13.78	-31.68	-20.31
		$AD_1^{(med)}$	22.24	15.30	29.48	19.70	23.71	26.12
3%	MLE	$D_1^{(med)}$	-301.24	-285.40	-427.23	-412.08	-494.90	-431.14
		$AD_1^{(med)}$	179.03	163.67	182.41	169.30	296.40	199.05
	King	$D_1^{(med)}$	71.22	64.73	83.08	79.99	156.95	141.05
		$AD_1^{(med)}$	59.20	54.28	76.20	59.95	79.23	65.40
	Firth	$D_1^{(med)}$	-9.11	-8.10	-13.99	-13.67	-16.76	-14.19
		$AD_1^{(med)}$	13.11	13.10	14.69	13.29	17.40	16.92
5%	MLE	$D_1^{(med)}$	-68.41	-66.04	-71.92	-68.74	-75.16	-72.28
		$AD_1^{(med)}$	66.21	66.29	78.17	67.89	84.39	73.16
	King	$D_1^{(med)}$	21.54	20.58	28.51	22.66	30.01	23.97
		$AD_1^{(med)}$	25.73	23.65	31.26	26.91	46.03	37.12
	Firth	$D_1^{(med)}$	-8.05	-7.44	-9.17	-8.75	-9.79	-9.48
		$AD_1^{(med)}$	11.64	11.34	12.59	11.62	13.94	11.62
10%	MLE	$D_1^{(med)}$	-35.21	-30.91	-36.81	-31.02	-51.73	-31.42
		$AD_1^{(med)}$	10.59	9.33	11.59	10.30	15.51	13.69
	King	$D_1^{(med)}$	13.84	12.68	15.16	13.25	16.48	13.29
		$AD_1^{(med)}$	7.78	7.20	11.20	7.98	24.40	19.20
	Firth	$D_1^{(med)}$	-0.93	-0.89	-1.86	-1.44	-5.12	-4.24
		$AD_1^{(med)}$	5.50	3.00	3.97	3.46	4.10	3.69

낸다. 'Complete case'는 결측치가 존재하지 않은 자료로 분석한 결과를 말하고 'Imputation'은 결측치를 대체값으로 채워 넣은 자료로 분석한 결과를 말한다. 분석 방법에서 'MLE'는 기존의 로지스틱 회귀분석 결과를 의미하고 'King'은 King, Zeng(2001)이 제시한 방법, 'Firth'는 Firth(1993)가 제시한 방법을 말한다. 각 β_j에 해당하는 추정치 $D_j^{(med)}$와 $AD_j^{(med)}$에 대해 반복적으로 비슷한 결과가 나왔으므로 논문에서는 j가 0,1일 경우만 제시하였다(Table 2, Table 3 참조).

$D_j^{(med)}$, $j=0,1$에 대한 분석결과는 Table 2와 Table 3에서 공통적으로 세 가지 특징을 가지고 있다. 먼저, Firth방법이 회귀계수 추정치 차이가 가장 적게 나타났고 MLE방법에서 차이가 가장 크게 나타났다. 또한, 파산발생률이 커짐에 따라서 세 방법에 차이가 줄어들었다. King방법은 기존의 로지스틱 회귀모형 추정치를 구한 후 보정을 하는 것이므로, 기존의 로지스틱 회귀모형 추정치가 발산하면 King방법의 추정치도 발산하게 된다. Firth방법은 다른 방법에 비해 파산발생률에 크게 영향을 받지 않아 희소사건의 분석에서 가장 신뢰할 수 있는 결과를 제시한다고 볼 수 있다. King의 방법이 MLE보다 좋은 결과를 보이지만 Firth보다는 성능이 떨어짐을 모의실험 결과를 통해 확인할

수 있다. 둘째, 결측률이 10%에서 50%로 갈수록 회귀계수 추정치의 차이가 커지고 있음을 보여준다. 이는 결측률이 높을수록 대체값에 포함되는 불확실성도 커지므로 당연할 결과라고 할 수 있다. 셋째, 대체값으로 분석한 결과인 'Imputation'값이 결측을 제외한 자료에서 분석한 결과인 'Complete case'보다 모의자료를 이용한 추정치 중위수와 전체자료를 이용한 추정치의 차이가 더 작음을 확인할 수 있다. 대체값이 적절하게 채워져 희소사건의 분석 시 좀 더 좋은 결론을 이끈다는 것을 의미한다. 자료의 파산발생률이 증가할수록(희소의 정도가 약해질수록), $AD_j^{(med)}$, $j=0,1$값이 줄어들었다. 또한 MLE, King, Firth 순으로 $AD_j^{(med)}$값이 줄어들었다.

6. 결론

신용평가 자료에서는 흔히 종속변수가 희소사건이며 자료에 결측이 존재한다. 일반적인 회귀모형 추정방법을 사용할 경우 파산발생확률이 과소 추정되는 경우가 발생할 수 있다. 또한 결측치가 존재하면 추정치에 추가적인 편향이 발생하고 파산여부 예측값 추정치의 분산이 증가하게 된다. 본 연구에서는 신용평가 자료를 모의생성하고 두 가지 결측치 처리방법과 세 가지 로지스틱 회귀 추정방법에 따른 결과를 비교하였다. FCS 알고리즘(Van Buuren, 2007)을 이용한 결측치 보정과 Firth(1993) 방법을 이용한 희소사건 편의 보정을 적용한 경우 가장 좋은 결과를 얻을 수 있었다. 특히 Firth(1993) 방법은 결측률과 희소사건의 정도에 민감하게 반응하지 않고 안정적인(robust) 결과를 보임을 확인할 수 있었다.

미국의 서브프라임 모기지 사태를 시작으로 확산된 2008년 금융위기 발생 이후 신용평가의 중요성이 더욱 부각되었다. 신용평가는 대출자의 신용위험에 대한 정보를 대부자에게 제공함으로써 대출자의 채무불이행시 대부자가 부담하는 손실을 줄일 수 있다. 신용평가 자료의 특징을 고려하지 않고 파산여부예측 모형을 개발한다면, 대부자는 대출자의 채무불이행 확률을 낮게 추정하여 변제능력이 부족한 대출자에게 카드발급이나 담보대출과 같은 금융상품을 판매할 가능성이 높다. 그 결과 변제능력이 부족한 기업이나 개인의 대출자금을 이용한 지출이 증가되어 파산율이 올라가게 된다. 또한 대부자 측면에서는 대출자의 채무불이행으로 인한 자금부족으로 경영 정상화를 기대하기 어려운 상황이 발생할 수 있다. 따라서 본 연구에서 사용한 방법들을 토대로 파산여부예측 모형을 개발한다면, 대출자의 채무불이행 확률을 낮게 추정하는 오류를 범하지 않으면서 파산가능 예측치의 정밀도를 높일 수 있을 것으로 기대된다.

References

Aalto University (2017). *Applications of Machine Learning Group - Datasets*. [online] Available at: http://research.ics.aalto.fi/eiml/datasets.shtml [Accessed on 5 Feb. 2017].

Buuren, S., Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R, *Journal of Statistical Software*, 45(3), 1-67.

Casella, G., George, E. I. (1992). Explaining the Gibbs sampler, *The American Statistician,* 46(3), 167-174.

Firth, D. (1993). Bias reduction of maximum likelihood estimates, *Biometrika*, 80(1), 27-38.

Gelman, A., Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models,* Cambridge University Press, 529-543.

Hand, D. J., Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review, *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 160(3), 523-541.

Heinze, G., Puhr, R. (2010). Bias reduced and separation proof conditional logistic regression with small or sparse data sets, *Statistics in Medicine,* 29(7-8), 770-777.

Heinze, G., Schemper, M. (2002). A solution to the problem of separation in logistic regression, *Statistics in Medicine,* 21(16), 2409-2419.

Jamshidian, M., Jennrich, R. I. (2000). Standard errors for EM estimation, *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 62(2), 257-270.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems, In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences,* 186(1007), 453-461.

Kim, H., Kang, C., Kim, T. (2012). Churn analysis on H discount stores using hierarchical logistic regression model, *Journal of the Korean Data Analysis Society*, 14(1), 117-124. (in Korean).

King, G., Zeng, L. (2001). Logistic regression in rare events data, *Political Analysis,* 9(2), 137-163.

KOSIS (2017). *Statistics Korea, Economically Active Population Survey* [online] Available at: http://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1DA7001&vw_cd=MT_ETITLE&list_id=&scrId=&seqNo=&language=en&obj_var_id=&itm_id=&co. [Accessed on 5 Feb. 2017].

Kwak, O., Kang, C., Choi, S., Kim, K. K. (2015). Development of personal credit scoring models by grouped customers, *Journal of the Korean Data Analysis Society,* 17(6), 3003-3014. (in Korean).

Lee, B. E. (2017). *Rare bankruptcy event prediction with missing data,* Unpublished Master's Thesis, Dongguk University, Seoul. (in Korean).

Lee, K. J., Carlin, J. B. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation, *American Journal of Epidemiology*, 171(5), 624-632.

Lee, Y., Song, J. (2007). Reject inference of a credit scoring model for data including binary characteristic variables, *Journal of the Korean Data Analysis Society,* 9(6), 2743-2754. (in Korean).

Little, J., Rubin, D. B. (2002). *Statistical Analysis with Missing Data,* Hoboken, NJ: Wiley, J., Sons. R.

McCullagh, P., Nelder, J. A. (1989). *Generalized linear models,* CRC press, 37.

Meng, X. L., Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm, *Journal of the American Statistical Association,* 86(416), 899-909.

OECD (2017). *Household accounts - Household debt - OECD Data*. [online] Available at: https://data.oecd.org/hha/household-debt.htm [Accessed on 5 Feb. 2017].

Park, J., Bang, S. (2015). Logistic regression with sampling techniques for the classification of imbalanced data, *Journal of the Korean Data Analysis Society,* 17(4B), 1877-1888. (in Korean).

Pastor, J. B. N. (2003). Methods for the analysis of explanatory linear regression models with missing data not at random, *Quality & Quantity*, 37(4), 363-376.

Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys,* John Wiley & Sons, 81.

Rubin, B. (1976). Inference and missing data, *Biometrika,* 63(3), 581-592.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification, *Statistical Methods in Medical Research,* 16(3), 219-242.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G. M., Rubin, D. B. (2006). Fully conditional specification in multivariate imputation, *Journal of Statistical Computation and Simulation,* 76(12), 1049-1064.

Yu, Q., Lendasse, A., Séverin, E. (2009). Ensemble KNNs for bankruptcy prediction, In *Computing in Economics and Finance 09, 15th International Conference: Computing in Economics and Finance,* 1-9.

Rare Bankruptcy Event Prediction with Missing Data

Bo Eun Lee[1], *Yong Sung Joo*[2], *Hyung Jun Jung*[3]

Abstract

Predicting bankruptcy has played an important role in credit evaluation. In particular, it is an indicator that has been used mainly in making decisions on bank loans. Bankruptcy data have two characteristics that require attention in statistical analysis. First, bankruptcy, a response variable of data, is a rare event. When the plain maximum likelihood estimation method is applied to the logistic regression model which predicts the probability of a rare event, the probability of occurrence tends to be underestimated due to bias of the regression coefficient estimate. Then, the lender can underestimate the possibility of bankruptcy and suffer losses due to lack of payment. Second, there may be missing value in the bankruptcy data. In the case of using only the complete case data, the accuracy of the model decreases. In addition, bias of estimate may arise in the model estimate. In this paper, we verify the effectiveness of the bias correction method of the bankruptcy prediction model through simulation studies.

Keywords : credit evaluation, rare event, logistic regression, imputation, bias correction.

[1]Graduate Student, Dep. of Statistics, Dongguk University, Seoul 04620, Korea.
E-mail : leebe9112@dongguk.edu
[2]Professor, Dep. of Statistics, Dongguk University, Seoul 04620, Korea.
E-mail : yongsungjoo@dongguk.edu
[3](Corresponding Author) Professor, Dep. of Taxation, Changwon University, Changwon 51140, Korea.
E-mail : jeon266@changwon.ac.kr
[Received 20 January 2017; Revised 7 February 2017, 17 February 2017; Accepted 20 February 2017]