응용통계연구 제 19권 3호, 2006년, pp. 407–419

이단계 일반화 선형모형을 이용한 은행 고객의 연체성향 분석*

오만숙[1] 오현탁[2] 이영미[3]

요 약

본 연구에서는 최근 몇 년 동안 한국에서 사회적으로 큰 문제가 되고 있는 카드이용 고객의 연체 방지를 위해 기존 금융권을 이용하는 개인에 대한 연체 성향 분석을 수행한다. 연체 성향 분석의 대상은 현재 한국의 한 특정 은행 고객으로 하였으며, 여러 연체 가운데 신용카드 연체를 중심으로 연구하였다. 연체 성향과 요인이 분석되면 기존 은행 고객에 대한 향후 부실의 정도를 예측할 수 있으며 또한 미래의 고객에 대해서도 신용평가 시스템을 만들 수 있을 것이다. 연체 성향 분석을 위한 통계적 방법으로 연체 유/무에 대한 로지스틱 회귀모형을 적용하였고 연체가 있을시 연체금액에 대한 일반화 선형모형을 적용하여 자료를 적합한 후, 유의한 설명변수들을 선정하여 반응변수와 설명변수들의 관계를 설명해 보았다. 분석 결과 연령, 건당 현금서비스 평균금액, 타사 현금금액, 수신잔액, 순수익이 연체 유무와 연체금액에 중요한 영향을 미치는 설명변수들임을 알 수 있었다.

주요용어: 신용평가시스템, 로지스틱 회귀모형, 감마분포, 변수선택

1. 서론

최근 몇 년간 가계의 신용대출이 크게 늘어나면서 사회 곳곳에서 부실화에 대한 우려의 목소리가 높아지고 있다. 2002년 9월말 우리나라의 가계 신용 잔액은 전 분기 6월말 보다 6.7% 늘어난 424.3조원에 이르렀고, 2003년 9월 중 은행 가계대출은 전월과 같은 3.2조원 증가하였다(한국은행, 2003). 1996년부터 2003년 말까지 가계 신용 대출 증감 추이를 살펴보면 그림 1.1과 같다. 그림 1.1에서 가계대출금액은 가계에 대한 대출, 현금서비스 및 카드론으로, 개인사업자금은 제외하되 주택자금은 포함한 것이고, 예금은행은 일반은행, 특수은행, 산업은행을 포괄한 대출금액이다. 또한 여신전문기관은 신용카드회사 및 할부금융회사를 포괄한 대출금액이다. 그림 1.1을 보면 1996년 1/4분기와 2/4분기 동안에 261조 8,526억원이었던 가계대출금액이 2003년 1/4분기와 2/4분기 동안에 799조 3,191억원으로 3배 이상으로 증가하였다. 이렇게 가계대출이 급증하게 된 원인은 다음과 같다. 먼저 저금리 정책

* 이 연구는 한국학술진흥재단의 연구비지원으로 수행되었음 (KRF-2002-070-C00017).

1) (120-750) 서울시 서대문구 대현동 21, 이화여자대학교 통계학과, 교수
E-mail: msoh@mm.ewha.ac.kr.
2) (561-765) 전주시 덕진구 덕진동 1가 664-14, 전북대학교 상경대학, 교수
E-mail: oht@chonbuk.ac.kr
3) (120-750) 서울시 서대문구 대현동 21, 이화여자대학교 통계학과, 대학원생
E-mail: loveadv@daum.net

으로 금융기관의 자금은 풍부해진 반면 경기 둔화에 따른 투자 부진과 부채 감축 노력으로 기업 자금 수요는 작아짐에 따라 생긴 유휴 자금을 금융기관이 수익성 높은 소매 금융 확대에 이용하고 있다는 공급 측면을 들 수 있다. 그리고 수요 측면에서는 가계가 상대적으로 문턱이 낮아진 금융 기관으로부터 저리 자금을 빌릴 수 있게 됨에 따라 소비 및 주택 관련 자금 지출을 늘렸기 때문이다(신수일, 2002).

이러한 가계여신의 증가와 맞물려 개인 신용 불량자의 양산은 2003년에도 두드러지고 있다. 전반적으로 가계 여신의 증가는 가계 신용 총계의 증가와 맞물려 있으며, 이러한 가계 여신의 증가 등으로 신용 불량자는 계속해서 증가하였으며, 정부의 잇따른 구제 조치에도 불구하고 증가세는 계속되었다. 특히 청년 신용 불량과 더불어 급속한 증가를 보였는데, 이는 경제적으로 많은 부담을 주었다. 이러한 신용 불량자 양산의 문제를 해결하기 위해서는 금융기관에서 신용 평가 시스템을 구축하여 개인이나 기업의 여신을 관리한다면 신용도의 기초가 될 수 있는 여러 대출금의 연체 방지에 많은 도움이 될 것이다. 최근 신용 평가 시스템의 중요성을 인식함에 따라 여러 기관에서 시스템의 개발 및 보완 작업이 한창 이루어지고 있으나 이의 발전은 아직 미약한 편이다.

흥미있는 것은 신용불량자의 대부분이 신용카드 관련 신용불량자라는 점이다. 그림 1.2는 2000년 이후의 전체 신용 불량자와 그 중 신용카드 관련 신용 불량자에 대한 증가 추세를 나타내는 그래프이다. 그림을 보면 신용불량자의 과반수 이상이 신용카드관련으로 비율이 커지는 추세이며 2003년 7월말에는 신용카드 관련 신용불량자가 207만을 기록하여 전체 신용불량자의 62%를 차지하고 있다(한국경제연구원,2003). 따라서 신용카드관련 신용불량자의 연체를 분석하는 것은 신용평가 시스템 구축의 첫번째 단계가 될 수 있을 것이다.

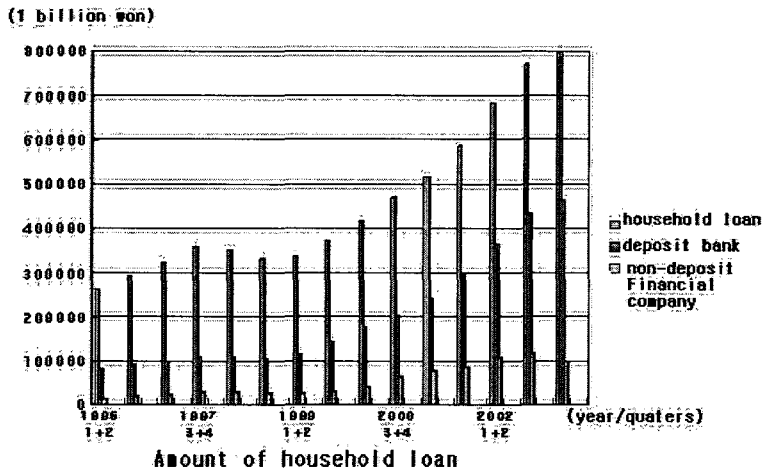

그림 1.1: 가계 신용대출 증감

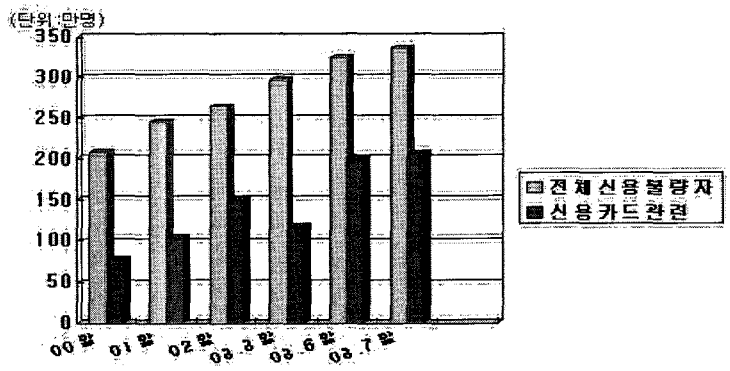

그림 1.2: 연도별 신용불량자 현황

신용카드 연체자료의 특징은 아주 소수를 제외한 다수의 카드 소지자들이 연체하지 않고 제때에 대금을 지불하지만 소수의 연체자들이 연체하는 금액이 매우 크다는 것이다. 따라서 연체자료의 전형적인 모양은 0에서 매우 높은 빈도를 갖고 오른쪽으로 긴 꼬리를 갖는 이산과 연속의 합성분포를 가진다. 이렇듯 관측치가 다수의 0과 소수의 0이 아닌 값으로 구성되어 있는 자료를 영과잉 또는 영팽창 (zero-inflated) 자료라 하는데 이는 보험이나 회계감사, 신뢰도, 계량경제학, 환경과학의 분야에서는 물론, 기상학, 농업, 의학, 제조업 등의 다양한 분야에서 볼 수 있다(Cameron and Trivedi, 1986; Dietz and Bohning, 2000; Hall, 2000; Kvanli et al, 1998; Lambert, 1992). 영과잉 자료는 관측치가 0에 쏠리는 비대칭성 때문에 정규모형이나 우도방법을 사용하여 분석하는데 어려움이 따른다. 연체 자료가 갖는 이런 문제들을 해결하기 위해 우리는 이단계 모형을 제안한다. 첫번째 단계에서는 로짓 모형을 사용하여 연체유무를 판별한다. 다음 단계에서는 연체가 있을 경우 연체금액에 대하여 오른 쪽으로 긴 꼬리를 갖는 연속분포를 사용하여 연체금액을 예측 또는 추정하고자 한다. 이 이단계 모형은 자료의 특징을 잘 설명할 뿐 아니라 어떤 사람이 연체할 확률과 연체시 연체금액을 예측할 수 있게 해준다.

각 단계마다 일반 선형모형을 사용하여 종속변수와 설명변수를 연결시켜주며 변수선택을 통해 연체에 유의한 영향을 미치는 설명변수를 찾아낸다(McCullagh and Nelder, 1989). 이렇게 얻어진 유의한 설명변수들을 가진 이단계 모형이 최종적으로 선택되며 이를 사용하여 연체를 예측하고 신용평가 시스템의 기반을 구축한다. 이 논문에서 우리는 신용카드 현금서비스를 반응변수로 한 경우만을 분석하나, 현금서비스 연체금액 이외에 신용카드와 관련하여 카드론 연체금액, 일시불 연체금액, 할부연체금액 등을 타깃 변수로 하여 마찬가지로 분석을 할 수 있을 것이다.

신용카드 연체 및 신용 불량자 분석은 Lee(1997), Lin et. al.(2001), 권오준(1997), 이혜승과 서용무(2003) 등이 있다. 그러나 이들은 모두 데이타 마이닝 기법을 사용하여 고객을

분류하는 것으로 본 논문에서와 같이 통계적 모형을 설정하여 분석하는 것과는 차이가 있다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 한국의 A 은행 고객에서 얻어진 자료를 설명한다. 분석 전의 자료정리와 자료의 특징을 간단히 기술한다. 3절에서는 영과잉과 오른쪽으로 치우친 자료의 특징을 반영하는 이단계 모형을 제안하고 이 모형을 사용한 통계분석 결과를 제시한다. 마지막 4절은 결론에 할애하였다.

2. 자료의 정리

본 연구의 분석 대상 자료는 한국 내 A은행의 개인여신 고객 프로파일을 근거로 한다. 이 고객 자료 중에서 본 연구에서 중점을 둔 사항은 고객이 보유한 신용카드 중에서 결제금액이 A은행 계좌에서 인출되는 고객 자료에 대한 분석이다. 데이터가 보유한 개인 신용변수는 71개로 구성되어 있으며, 조사 기간은 2001년 3월부터 2002년 2월까지 1년간이고, 각각의 레코드에 포함된 금액과 건수는 1년 누적 건수와 금액이다.

본 연구에서 사용하는 데이터는 총 71개의 변수로 구성되어 있는데 이 변수들을 모두 모형에 포함시키기에는 변수의 수가 많으므로 사전에 중요성이 덜하다고 생각되는 변수들은 제거하였다. 먼저 변수들 중에 수수료와 관련된 변수, 해외사용과 관련된 변수, 신용카드 부가 기능으로 생긴 변수는 사전에 고려대상에서 제외하였다. 관련된 변수들 중 금액과 건수로 분리되어 있는 변수는 건수별 금액으로 다시 환산하여 평균액수(금액/건수)를 분석에 사용하였으며 모든 금액과 관련된 변수의 단위는 100만원으로 설정하였다. 또한 금액변수들에 대한 관측치 중 음의 값은 오타로 간주하여 모두 사전에 삭제하였다.

이렇게 얻어진 설명변수들간의 상관관계를 살펴본 결과, 높은 상관관계를 갖는 변수들이 다수 발견되었다. 이는 설명변수들 간에 중복성이 있는 것으로 판단되어 상관계수가 0.5 이상인 변수들은 사전에 변환하거나 일부를 제거하여 다중공선성의 문제를 해결하였다. 표 2.1은 최종적으로 분석에 사용된 13개의 설명변수들이며 각 종속변수와 설명변수에 대한 14,045개의 자료가 분석에 사용되었다. 위 변수 중 타사 신판금액은 타사의 신용판매 금액을 나타내고 타사 현금금액은 타사의 현금서비스사용금액을 나타낸다.

표 2.1: 모형분석에 사용된 변수

연령	당행카드 소지수
수신잔액	순수익
신탁잔액	할부평균(금액/건수)
여신 잔액	카드론 약정액
일시불 평균 (금액/건수)	현금서비스 평균(금액/건수)
타사현금 금액	타사카드 소지수
타사 신판금액	

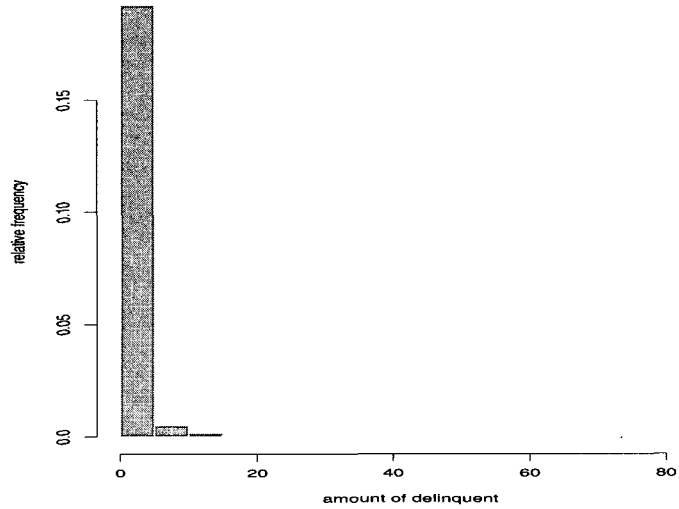

그림 2.1: 연체금액 히스토그램

3. 통계적 분석

앞 절에서 언급한 바와 같이 타깃변수는 '현금서비스 연체금액'으로 정하였으며, 이 변수도 금액을 나타내는 다른 변수들과 마찬가지로 100만원 단위로 바꾸었다. 그림 3.1은 타깃변수의 히스토그램이다. 이 그림을 살펴보면, 타깃변수의 관측치가 매우 많은 영의 값(zero)과 소수의 영이 아닌 값(nonzero)으로 되어 있다. 이는 전체 은행의 카드고객 중 연체가 발생하는 비율이 극히 낮으므로 현금서비스 연체 금액이 영의 값(zero)을 가지는 경우가 대부분이고 소수의 관측치만이 연체가 발생하는 경우에 해당하여 양의 값을 가지기 때문이다.

자료의 영과잉 특성을 반영하기 위하여, 우리는 종속변수인 현금서비스 연체금액에 대하여 이단계 모형을 제안한다. 먼저 타깃변수 '현금서비스연체금액'이 영보다 클 경우는 1값을 갖고, 연체금액이 없을 경우에는 0값을 갖는 '현금서비스 연체금액 유/무'라는 새로운 이항변수를 생성하여, 이를 반응변수로 하는 로지스틱 회귀 모형을 구축한다. 이 모형으로 미래의 잠재적인 고객에 대한 연체 유/무를 예측할 수 있을 것이다. 다음, 연체가 있을 경우 우리의 관심은 '연체금액' 자체가 된다. 이 경우에 연체금액은 양의 값을 갖는 연속자료이므로 연체금액에 대한 연속 모형을 사용하여 자료를 적합시킨다. 이 두 단계를 합성하여 연체확률과 동시에 연체시 연체금액을 예측할 수 있다.

이 절에서는 실제로 데이터를 모형에 적합해본다. 은행 고객들의 연체에 영향을 미치는 요인들과 종속변수와의 관계구조를 알아보기 위하여, 단계적으로 로지스틱 회귀 모형과 일반화 선형 모형을 사용하여 모형별로 유의한 변수 선정과 각 모형의 적합도 검정이 수

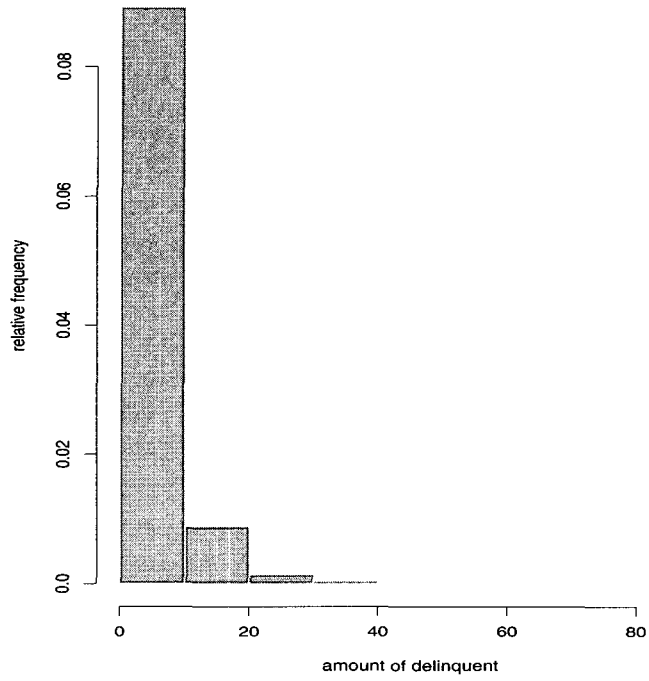

그림 3.1: 연체가 있는 경우 연체금액의 히스토그램

행된다. 모든 데이터 처리와 분석은 The SAS System for Windows V8을 사용하였다(SAS, 2000).

로지스틱 회귀모형을 적합하기 전에, 반응변수 "현금서비스 연체 유/무"의 비율을 계산해보니 연체금액이 있는 고객들이 전체의 11.6%였다. 13개의 설명변수 중 현금서비스 연체 유/무에 영향을 미치는 변수들을 선택하기 위해 먼저 단계별 변수 선정법(stepwise method)을 이용하였다. 진입기준(signigicant level of entry)으로는 0.15, 제거 기준(significant level of stay)도 0.15를 사용하였다. 이 방법으로 선택된 변수는 '연령', '타사신판금액', '타사현금금액', '수신잔액', '현금서비스금액평균(금액/건수)', '순수익' , 그리고 '당행카드소지수'이다. 이들 변수의 회귀계수 추정치와 오차, 유의확률이 표 3.1에 정리되어 있다. 선정된 설명변수가 모두 약 1% 유의수준으로 유의함을 알 수 있다.

이 방법으로 추정된 로지스틱 회귀 모형식은 다음과 같다.

$$\begin{aligned}\log(\frac{\hat{p}}{1-\hat{p}}) &= 1.4733 + 0.0088 \text{ 연령} - 0.0128 \text{ 타사신판금액} + 0.0065 \text{ 타사현금금액} \\ &\quad +0.0600 \text{ 수신잔액} + 0.2855 \text{ 평균현금서비스금액} - 2.6995 \text{ 순수익} \\ &\quad +0.3291 \text{ 당행카드소지수}.\end{aligned}$$

표 3.1: 로지스틱 모형의 회귀계수 추정치

Variable	df	Estimate	SE	Chi-Square	p-value
Intercept	1	1.4733	0.1795	67.3363	< 0.0001
연령	1	0.0088	0.0036	5.8764	0.0153
타사신판금액	1	-0.0128	0.0024	28.5813	< 0.0001
타사현금금액	1	0.0065	0.0010	44.2278	< 0.0001
수신잔액	1	0.0600	0.0114	27.7927	< 0.0001
현금서비스 평균금액	1	0.2855	0.0895	10.1763	0.0014
순수익	1	-2.6995	0.2064	171.0968	< 0.0001
당행카드소지수	1	0.3291	0.1208	0.7416	0.0065

여기에서 $\hat{p}$ 는 연체확률의 추정치이다.

위의 회귀 모형식을 살펴보면 연령, 수신잔액, 평균 현금서비스금액, 당행카드소지 수의 계수는 양의 값을 갖는 반면, 타사신판금액,타사현금금액 ,순수익의 계수는 음의 값을 갖는다. 이는 연령이 높을수록, 타사신판금액이 낮을수록, 타사현금금액이 낮을수록, 수신잔액이 높을수록, 평균 현금서비스금액이 높을수록, 순수익이 낮을수록, 당행카드소지수가 높을수록 '현금서비스 연체'가 있을 확률이 높다고 추정할 수 있음을 의미한다. 특히 순수익은 '현금서비스 연체'가 없음에 큰 영향을 주고 건당 평균현금서비스금액과 당행카드소지수는 '현금서비스 연체' 있음에 큰 영향을 줌을 알 수 있다. 이와같은 결과는 표 3.2에 주어진 회귀계수에 대한 오즈비(Odds Ratio)에서도 마찬가지로 관찰될 수 있다.

표 3.2: 로지스틱 모형으로부터 추정된 연체 유무의 로그오즈비

Variable	오즈비	95% Wald C.I	
연령	1.009	1.002	1.016
타사신판금액	0.987	0.983	0.992
타사현금금액	1.007	1.005	1.008
수신잔액	1.062	1.038	1.086
현금서비스 평균금액	1.330	1.116	1.585
순수익	0.067	0.045	0.101
당행카드소지수	1.390	1.097	1.761

다음으로 표 3.3는 로지스틱 회귀 분석의 정오 분류표이다. 정오 분류표는 목표변수의 실제 범주와 모형에 의해 예측된 분류범주 사이의 관계를 나타내는 표이다. 표를 통해 오분류율, 정확도, 민감도, 특이도를 얻어 모형을 비교할 수가 있다.

표 3.3에 있는 정확도(Accuracy), 민감계수(Sensitivity), 특이계수(Specificity), 오분류율(Error rate)의 의미를 살펴보면 다음과 같다. 민감계수는 실제 불량인 개인을 불량으로 분류하는 지수로, 71.9%로 불량집단을 예측하는 것을 알 수 있다. 특이계수는 실제 우량인 개인을 우량으로 분류하는 것으로서 이 회귀모형의 경우, 52.5%로 예측하고 있음을 알 수

표 3.3: 연체 유무에 따른 고객의 분류

		True			예측력	
		연체 무	연체 유	총계	민감계수	71.9%
추정	연체 무	5927	595	6522	특이계수	52.5 %
	연체 유	657	2313	2972	분류정확도	69.3%
	총	6584	2910	9494	양의 예측값	90.9%

있다. 분류정확도란 전체 사례수를 실제 우량을 우량으로 분류한 수와 실제 불량을 불량으로 분류한 수를 합한 것으로 나눈 것을 의미한다. 이 모형의 경우, 69.3%의 정확도를 보인다. 오분류율은 이 정확도의 반대의 개념이라 할 수 있다. 또한 실제 우량인 집단을 우량으로 분류한 수를 우량으로 분류한 수로 나누면 그것이 양의 예측값이 되는데, 이 모형에서는 양의 예측값이 90.9%로 나타났다. 실제 데이터를 사용하여 모형을 구축했다는 점을 감안하여 볼 때 위의 수치들을 통하여 예측 결과가 우수하다고 말할 수 있다.

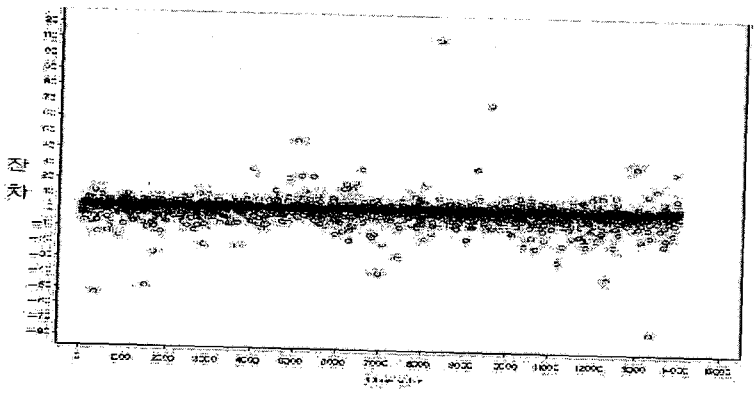

그림 3.2: 감마 모형의 잔차그림

지금까지 로지스틱 회귀모형을 통하여, 현금서비스 이용고객의 연체 유/무를 예측하는 모형을 추정하였다. 이제 연체를 행하는 고객에 대해 연체금액을 예측하는 모형을 구축해 보고자 한다. 이 모형에서는 타깃변수를 연속된 값 그대로 일반화 선형 모형(Generalized Linear Model)을 적합해 보았다. 모형을 적합하기에 앞서, 타깃변수인 "현금서비스연체금액" 변수의 분포 중 연체금액이 있는 경우의 분포를 살펴보고자 히스토그램을 그려보았다. 그림 3.2는 현금서비스 연체금액을 백만원 단위로 환산한 히스토그램이다. Y축은 변수 값의 백분율이다. 히스토그램을 보면, 현금서비스 연체금액의 분포가 정규분포와 다름을 알 수 있다. 따라서 본 연구에서는 일반화 선형모형을 적용하기로 하고 '현금서비스 연체금액

'의 분포를 보아, 감마분포를 적용하여 적합해 보았다. 또한 일반화 선형모형에서는 가법성을 연관함수(Link function)을 통해 성립한다고 하는데, 본 연구의 분석에서는 연관함수를 로그함수로 정하여 적합하였다. 분포함수로 가정한 로그 감마 모형의 확률밀도함수는 다음과 같다.

$$f(y) = \frac{1}{\Gamma(\nu)y}\left(\frac{y\nu}{\mu}\right)^{\nu} exp\left(-\frac{y\nu}{\mu}\right), \quad 0 < y < \infty.$$

또한 연관함수로 정한 $g(\mu) = log\mu$를 통해 성립되는 가법성을 표현하면 다음과 같다.

$$\eta = \log(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

SAS GENMOD PROCEDURE를 사용하여 로그 감마 일반화 선형 모형의 적합한 결과는 다음과 같다. 1단계에서와 같이 변수선택과정을 거쳐 6개의 유의한 설명변수를 선택하였다.

표 3.4: 감마 모형의 계수 추정치

Variable	Estimate	SE	Wald 95% C.I.		Chi-Square	Pr>Chi-Sq
Intercept	0.5230	0.1357	0.2570	0.7890	14.85	0.0001
연령	0.0058	0.0028	0.0003	0.0113	4.20	0.0405
타사카드소지수	0.0364	0.0140	0.0089	0.0639	6.74	0.0094
타사현금금액	-0.0021	0.0009	-0.0038	-0.0063	5.10	0.0239
수신잔액	-0.0140	0.0039	-0.0217	-0.0063	12.76	0.0004
건당 현금서비스평균금액	0.3257	0.0957	0.1381	0.5134	11.58	0.0007
순수익	2.0147	0.1819	1.6582	2.3711	122.74	<0.0001

표 3.4는 반복모수추정방법(iterative parameter estimation process)으로 얻어진 계수의 추정결과를 나타낸다. 이로부터 추정된 연관함수는 다음과 같이 표현할 수 있다.

$$\begin{aligned}\eta = \log(\mu) \quad = \quad & 0.5230 + 0.0058 \text{ 연령} + 0.0364 \text{ 타사카드 소지수} - 0.0021 \text{ 타사현금금액} \\ & -0.0140 \text{ 수신잔액} + 0.3257 \text{ 건당 현금서비스 평균금액} + 2.0147 \text{ 순수익}.\end{aligned}$$

연령, 타사카드 소지수, 건당 현금서비스 평균금액, 순수익 등은 양의 계수를 갖는 반면 타사현금금액, 수신잔액은 음의 계수를 갖는다. 좀더 구체적으로, 연령, 타사카드 소지수, 건당 현금서비스 평균금액, 순수익이 한단위 증가할 때 우리는 연체금액이 각각 $0.58\%(e^{0.0058}-1)$, $3.7\%(e^{0.0364}-1)$, $38.5\%(e^{0.3257}-1)$, $649.85\%(e^{2.0147}-1)$ 증가할 것으로 기대한다. 그리고 타사현금금액, 수신잔액이 한 단위 증가할 때 마다 연체금액이 각각 $0.21\%(e^{0.0021}-1)$, $1.41\%(e^{0.014}-1)$ 감소할 것으로 기대한다.

다음의 표 3.5는 추정된 로그 감마 일반화 선형모형의 Type 3 분석을 위한 우도비 통계량을 나타낸다. Type 3 분석은 다른 변수가 이미 모형에 포함되어 있을 때, 해당 변수가 모형에 얼마나 유의한가를 검정하는 방법이다. 모든 변수가 매우 유의함을 알 수 있다.

표 3.5: 감마 모형의 타입 3 우도비 검정 통계량

Variable	DF	Chi-Square	Pr > Chi-Sq
연령	1	4.25	0.0392
타사카드소지수	1	6.81	0.0091
타사현금금액	1	4.98	0.0257
수신잔액	1	8.72	0.0031
건당 현금서비스평균금액	1	12.39	0.0004
순수익	1	132.73	< 0.0001

표 3.6: 감마 모형의 적합도 검정

Criterion	DF	Value	Value/DF
Deviance	1245	1048.8329	0.8424

표 3.6에 제시된 모형의 적합도 검정 결과를 보면 유의확률이 0.9로 제안된 로그감마 모형이 적합하다고 결론 짓는다. 그림 3.2 은 로그감마 모형의 잔차그림으로 관측된 연체금액과 추정된 연체금액의 차이를 보여준다. 몇 개의 이상치를 제외하고 거의 모든 잔차가 영을 중심으로 수평밴드 내에 모여 있으므로 주어진 모형이 적합함을 나타내고 있다.

첫 번째 단계와 두번째 단계의 분석결과를 종합하여 각 변수에 대응하는 계수의 음양을 표 3.7에 표시하였다. 빈 칸은 그 변수가 유의하지 않음을 나타낸다. 총 8개의 변수가 연체에 유의한 영향을 미치는데 그 중 연령, 건당 현금서비스 평균금액, 타사현금금액, 수신잔액, 순수익은 연체 유무와 연체금액 모두에 영향을 미치며 특히 연령, 건당 현금서비스 평균금액은 두 단계에서 모두 양의 영향을 미친다. 따라서 연령과 건당 현금서비스 평균금액이 높아질수록 연체 위험이 높아진다고 볼 수 있다.

표 3.7: 이단계 모형의 각 단계에서 유의한 설명변수의 계수의 부호

변수	로지스틱 모형 (1단계)	감마모형 (2단계)
연령	+	+
건당 현금서비스 평균금액	+	+
순수익	−	+
수신잔액	+	−
타사현금금액	+	−
당행카드소지수	+	
타사카드소지수		+
타사신판금액	−	

그런데 무엇 보다도 흥미있는 것은 타사현금금액,수신잔액, 순수익이 두 단계에서 다른 부호를 가진다는 것이다. 타사현금금액, 수신잔액은 연체유무에는 양의 영향을 미치지만

연체시 연체금액에 대해서는 음의 영향을 미친다. 따라서 타사현금금액, 수신잔액이 많은 고객은 그렇지 않은 고객에 비하여 연체할 가능성이 높으나 연체금액은 작다고 볼 수 있다. A 은행 혹은 다른 은행에서 현금 서비스를 많이 받는 고객은 경제적 통제력이 약하여 연체할 위험이 높다고 할 수 있다. 그러나 이들이 A은행에 연체하는 금액은 적은데 이는 필요한 금액을 A 은행과 다른 은행에서 분산하여 빌리기 때문으로 유추할 수 있다. 수신잔액은 여신잔액과 피어슨 상관계수가 0.4387로 매우 밀접히 연관되어 있는데 이로 인하여 수신잔액이 연체여부에 양의 영향을 주는 것으로 보인다. 그리고 연체금액에 대한 수신잔액의 양의 영향은 아마도 수신잔액이 많은 고객들이 수신잔액의 일부를 사용하여 연체금액을 갚을 수 있다고 생각하기 때문으로 보인다.

순수익은 수신잔액과 타사현금금액과 반대로 연체 여부에는 음의 영향을 미치나 연체금액에는 양의 영향을 미친다. 순수익이 많은 사람은 연체를 하지 않는 경향이 있으나 만약 연체를 할 경우 연체 금액이 높음을 의미한다. 순수익과 건당 현금서비스 평균금액의 관계를 보면 피어슨 상관계수가 0.4577로 순수익이 많은 사람은 현금서비스 액수가 많고 따라서 연체 금액이 많은 경향이 있다.

당행카드 소지수, 타사카드소지 수가 연체 여부와 연체금액에 양의 영향을 미치는 것은 상식적으로 예견할 수 있다. 타사현금금액은 연체 여부에는 음의 영향을 미친다. 이는 타사현금 서비스를 많이 받는 사람은 A은행이 주거래 은행이 아니거나 다른 은행에서 돈을 빌려 A은행의 연체를 갚기 때문이라 유추할 수 있다.

4. 요약 및 결론

본 연구는 A은행 고객 자료를 이용하여 고객의 연체 성향을 분석하였다. 특히 신용불량과 밀접한 관련이 있는 '현금서비스연체금액'을 중점으로 다루었다. 자료는 한국의 A은행의 고객자료를 이용하였고 중복성 등을 고려하여 13개의 설명변수들을 선택하였다.

자료는 0에서 높은 빈도를 갖고 오른쪽으로 긴 꼬리를 갖는 분포를 갖는다. 이러한 영과잉 비대칭 자료의 특성을 고려하여 이단계 일반화 선형 모형을 제안하였다. 1단계에서는 로지스틱 모형을 사용하여 연체 유/무를 예측하고 2단계에서는 연체가 있을시 감마모형을 사용하여 연체 금액을 예측하였다. 13개의 설명변수 중 단계별 변수선택 방법을 사용하여 유의한 설명변수들을 색출하였다. 적합도 검정과 다른 모형진단을 시행한 결과 제안된 모형이 자료를 잘 적합함을 알 수 있었다.

유의한 설명변수로 선택된 것은 연령, 타사현금금액, 수신잔액, 건당 현금서비스 평균금액, 순수익, 타사카드 소지수, 타사신판금액, 당행카드 소지 수 등의 8개 변수이다. 이 중 5개 변수(연령, 건당 현금서비스 평균금액, 타사현금금액, 수신잔액, 순수익)는 연체여부와 연체금액 모두에 유의한 영향을 미친다. 평균적으로 연령, 건당 현금서비스 평균금액이 증가할수록 연체 위험이 높아진다. 그러나 타사현금금액, 수신잔액, 순수익은 연체여부와 연체금액에 서로 다른 부호의 계수를 갖는데, 타사현금금액, 수신잔액은 연체여부에는 양의 영향을 미치나 연체금액에는 음의 영향을 미치고 순수익은 그 반대이다. 이 세 변수와 또 다른 의미있는 설명변수에 대한 좀더 자세한 조사가 향후 연구과제가 될 수 있을 것이다.

참고문헌

신수일 (2002). 연관 규칙과 분류 규칙을 이용한 은행 고객의 연체 성향 분석에 관한 연구. 석사학위논문, 서강대학교 대학원.

한국경제연구원 (2003). <경제전망과 정책과제 2003년9월> **13(3)**, 한국경제연구원.

한국은행 (2003). <2003년 8월 중 금융시장 동향>, 한국은행.

Cameron, A. C. and Trivedi, P. K. (1986). Econometric models based on count data: comparisons and applications of some estimators and tests, *Journal of Econometrics* **1**, 29-53.

Dietz, E. and Bohning, D. (2000). On estimation of the Poisson parameter in zero-modified Poisson models, *Computational Statistics and Data Analysis* **34**, 441-451.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study, *Biometrics* **56**, 1030-1039.

Kvanli, A. H., Shen, Y. K., and Deng, L. Y. (1998). Construction of confidence intervals for the mean of a population containing many zero values, *Journal of Business and Economics* **16**, 362-368.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to random defects in manufacturing, *Technometrics* **34**, 1-14.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models, Chapman and Hall: London.*

SAS (2000). *SAS System for Windows V.8*, SAS Institute Inc.

[2006년 5월 접수, 2006년 7월 채택]

Analysis of Household Overdue Loans by Using a Two-stage Generalized Linear Model*

Man-Suk Oh[1] Hyeon Tak Oh[2] Young-Mi Lee[3]

ABSTRACT

In this paper, we analyze household overdue loans in Korea which has been causing serious social and economical problems. We consider customers of Bank A in Korea and focus on overdue cash services which have been snowballing in the past few years. From analysis of overdue loans, one can predict possible delays for current customers as well as build a credit evaluation and risk management system for future customers. As a statistical analytical tool, we propose a two-stage Generalized Linear regression Model (GLM) which assumes a logistic model for presence/non-presence of overdue and a gamma model for the amount of overdue in the case of overdue. We perform goodness of fit test for the two-stage model and select significant explanatory variables in each stage of the model. It turns out that age, the amount of credit loans from other financial companies, the amount of cash service from other companies, debit balance, the average amount of cash service, and net profit are important explanatory variables relevant to overdue credit card cash service in Korea.

Keywords: Credit evaluation system; Logistic regression model; Gamma distribution; Variable selection.

* This research was supported by Grant from Korea Research Foundation KRF-2002-070-C00017.

1) Professor, Dept. of Statistics, Ewha Womans University, Sodaemoon Gu, Seoul 120-750, Korea
E-mail: msoh@ewha.ac.kr

2) Professor, School of Business, Chonbook University, Dukjin Gu, Jeonju 561-765, Korea
E-mail: oht@business.cbu.ac.kr

3) Graduate Student, Dept. of Statistics, Ewha Womans University, Sodaemoon Gu, Seoul 120-750, Korea
E-mail: loveadv@daum.net