신용거래점수모형에 대한 고찰
- 순서 로지스틱 회귀모형 -

정 동 빈*

Study on Credit Scoring Model
- Ordinal Logistic Regression Model -

< 목차 >

개 요
Ⅰ. 서 론
Ⅱ. 순서 로지스틱 회귀모형
Ⅲ. 사례연구 : 소비자 신용자료
Ⅳ. 결 론
참고문헌
ABSTRACT

개 요

신용거래점수는 각 개인의 신용위험수준을 반영하는 기준이다. 본 논문에서는 신용거래점수모형의 한 방법으로 순서 로지스틱 회귀모형을 소개하며, 이 모형을 소비자 신용자료에 적용시켜 그 응용도를 살펴보았다. 이 모형을 통해 사회전반에 복잡하게 얽혀있는 경제현상을 부분적으로 설명할 수 있으며, 더 나아가 각 개인의 신용거래에 대한 의사결정을 하는데 도움이 될 수 있다.

주제어 : 신용거래점수, 연결함수, 최대가능도 추정량, 순서 로지스틱 회귀모형, 비례승산모형

* 강릉대학교 사연과학대학 정보통계학교수
접수일자 : 2003 5-11 게재확정일자 : 2003-7-18

정 동 빈

I. 서 론

신용거래점수(credit scoring)는 지난 수 십 년에 걸쳐 경영학 분야에서 고려한 가장 성공적인 통계적인 기법으로, 실제자료에 근거하였기 때문에 각 개인의 신용가치를 질적으로 평가하는 방법보다 더 정확도 있는 것으로 여겨왔다. 물론 신용거래점수를 고려할 때 관련된 집단(예: 최근 채무자들)으로부터 표본을 추출하여 어떤 요인들이 각 개인의 가치에 가장 영향을 주는지를 분석한다. 이에 관련된 요인은 미결제된 부채, 최근 신용계좌의 수와 유형, 연령, 소득, 과거 내부상태, 직업, 거주형태 등이 될 수 있다.

이에 관련된 주된 통계적인 기법은 다중회귀(multiple regression)분석과 로지스틱 회귀(logistic regression)분석이며, 이들을 통해 신용거래점수와 발생한 손실 사이의 관계가 특정적인 변수들로 설명되는지 여부와 기존 설명변수들에 포함되지 않은 손실에 관한 새로운 정보를 신용거래점수에 추가시켜야 하는지 여부를 판가름할 수 있게 된다.

5개의 수준(해당 날짜에 갚음, 만기일 30일 이내 상환, 만기일 60일 이내 상환, 만기일 90일 이내 상환, 상환불능)이 있는 변수 신용계좌상태를 고려해보자. 이 경우 각 범주에 순서가 있다. 즉, 해당 날짜에 갚는 것은 만기일 30일 이내보다 신용상태가 좋고, 마찬가지로 만기일 60일 이내보다는 만기일 30일 이내가 좋다. 그러나 그 차이가 항상 동일하다고 할 수 없을 것이다. 예를 들면, 만기일 90일 이내 상환과 상환불능의 차이와 만기일 30일 이내 상환과 60일 이내 상환 사이의 차이는 비록 각각 근접하는 범주라 하더라도 서로 다를 것이다. 이러한 변수를 순서형 변수(ordinal variable)라고 할 때, 이 변수를 종속변수로 고려한 로지스틱 회귀모형을 순서 로지스틱 회귀모형이라고 한다. 이와 같이 5개의 수준이 있는 순서형 변수(신용계좌상태)를 종속변수로 고려한다면, 이에 관련된 신용거래모형들 중 순서 로지스틱 회귀모형을 사용할 수 있을 것이다.

본 논문은 신용거래점수에서 가장 현실성 있게 사용할 수 있는 통계적 기법들 중 하나인 순서 로지스틱 회귀모형을 실제 자료에 적용시켜 그 쓰임도와 응용도를 알아보는데 그 의의가 있다. 또한 순서 로지스틱 회귀모형을 구체적으로 적용하기 위해서 통계패키지 SPSS 프로시저인 PLUM을 사용하기로 한다. 2장에서는 신용거래점수모형으로 사용할 수 있는 순서 로지스틱 회귀모형과 이 모형을 적합시키는 이론적인 방법에 대해 소개한다. 3장에서는 실제 추출한 1000명의 소비자 신용자료를 순서 로지스틱 회귀모형에 적용시켜 적절한 모형을 찾는 과정과 모형예측에 대해 살펴보도록 한다. 끝으로 4장에서 결론을 내린다.

II. 순서 로지스틱 회귀모형

순서형 종속변수를 다룰 로지스틱 회귀모형으로 비례승산모형(proportional odds model)을 알아보기로 한다(Hosmer와 Lemeshow, 2000). 이 모형은 $y \le k$과 $y > k$의 확률을 비교하며, 다음과 같이 표현된다.

$$c_k(\boldsymbol{x}) = \ln\left[\frac{P(y\le k|\boldsymbol{x})}{P(y>k|\boldsymbol{x})}\right]$$

$$= \ln\left[\frac{\pi_0(\boldsymbol{x})+\pi_1(\boldsymbol{x})+\cdots+\pi_k(\boldsymbol{x})}{\pi_{k+1}(\boldsymbol{x})+\pi_{k+2}(\boldsymbol{x})+\cdots+\pi_K(\boldsymbol{x})}\right]$$

$$= \alpha_k + \boldsymbol{x}'\boldsymbol{\beta}, \quad k=0,1,\cdots,K-1 \tag{2.1}$$

여기에서, $(p+1)\times 1$의 공변량벡터 $\boldsymbol{x}$가 주어졌을 때 결과변수가 k일 확률을 $P(y=k|\boldsymbol{x}) = \pi_k(\boldsymbol{x})$, α_k는 k번째 로짓에 대한 절편항, $\boldsymbol{\beta}$는 독립변수 $\boldsymbol{x}$의 회귀계수를, $(K+1)$은 종속변수 수준의 수를 각각 나타낸다. $K=1$인 경우 식 (2.1)에 주어진 모형은 $y=0$ 대 $y=1$의 확률의 비율을 계산할 수 있다는 점에서 통상적인 로지스틱 회귀모형으로 단순화할 수 있음을 알 수 있다.

모형을 적합시키는 방법을 다음 두 단계를 통하여 살펴보기로 하자.

1단계

식 (2.1)을 이용하여 종속변수가 특정한 순서범주에 속할 확률을 아래와 같이 구할 수 있다.

$$\pi_0(\boldsymbol{x}) = P(y=0|\boldsymbol{x}) = \frac{\exp(\alpha_0+\boldsymbol{\beta}'\boldsymbol{x})}{1+\exp(\alpha_0+\boldsymbol{\beta}'\boldsymbol{x})}$$

$$\pi_1(\boldsymbol{x}) = P(y\le 1|\boldsymbol{x}) - \pi_0(\boldsymbol{x})$$

$$= \frac{\exp(\alpha_1+\boldsymbol{\beta}'\boldsymbol{x})}{1+\exp(\alpha_1+\boldsymbol{\beta}'\boldsymbol{x})} - \frac{\exp(\alpha_0+\boldsymbol{\beta}'\boldsymbol{x})}{1+\exp(\alpha_0+\boldsymbol{\beta}'\boldsymbol{x})}$$

$$\cdots$$

$$\pi_{K-1}(\boldsymbol{x}) = \frac{\exp(\alpha_{K-1}+\boldsymbol{\beta}'\boldsymbol{x})}{1+\exp(\alpha_{K-1}+\boldsymbol{\beta}'\boldsymbol{x})} - \frac{\exp(\alpha_{K-2}+\boldsymbol{\beta}'\boldsymbol{x})}{1+\exp(\alpha_{K-2}+\boldsymbol{\beta}'\boldsymbol{x})} \tag{2.2}$$

$$\pi_K(\boldsymbol{x}) = 1-\pi_0(\boldsymbol{x}) - \cdots - \pi_{K-1}(\boldsymbol{x}).$$

2단계

$(K+1)\times 1$ 벡터 $\boldsymbol{z} = (z_0,\cdots,z_K)'$는 $y=k$일 때 $z_k=1$, $j\neq k$에 대해 $z_j=0$으로 놓는다. 따라서 변수 $\boldsymbol{z}$에서 단지 한 성분의 값이 "1"이 된다. 이때 n개의 독립된 관측값

(y_i, x_i), $i=1, \cdots, n$에 대한 가능도함수는

$$l(\beta, \alpha) = \prod_{i=1}^{n} \left[\pi_0(x_i)^{z_{0i}} \pi_1(x_i)^{z_{1i}} \times \cdots \times \pi_K(x_i)^{z_{Ki}} \right]$$

이다. 여기에서, β는 p개의 독립변수의 기울기, α는 K개의 절편항들이다. 로그 가능도함수는

$$L(\beta, \alpha) = \sum_{i=1}^{n} \{ z_{0i} \ln[\pi_0(x_i)] + z_{1i} \ln[\pi_1(x_i)] + \cdots + z_{Ki} \ln[\pi_K(x_i)] \} \quad (2.3)$$

이다. 모수에 대한 최대가능도 추정량은 식 (2.3)을 미지의 모수들에 대해 편미분한 $(K+p)$개의 방정식을 "0"으로 놓고 $(\hat{\beta}, \hat{\alpha})$를 풀면 된다. $(\hat{\beta}, \hat{\alpha})$에 대해 이차 편미분을 하여 얻은 행렬에 음수를 곱하여 정보행렬을 얻고, 그 역행렬을 구하여 추정된 계수의 공분산행렬을 추정할 수 있다.

비례승산모형을 일반화한 일반화 선형모형(generalized linear model)을 사용하여 분석을 할 수 있다(McCullagh과 Nelder, 1989). 일반화 선형모형은 폭넓은 범위의 통계적 모형에 사용할 수 있는 매우 강력한 모형이며, 그 일반적인 형태는 다음과 같다.

$$\text{link}(\gamma_j) = \theta_j - [\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k] \quad (2.4)$$

여기에서, link는 연결함수(link function)

γ_j는 j번째 범주에 대한 누적확률

θ_j는 처음 j번째 범주에 대한 절편항

$\beta_1, \cdots, \beta_k$는 회귀계수

$x_1, \cdots, x_k$는 독립변수

k는 독립변수의 수이다.

여기에서 다음 주어진 몇 가지 사항들을 고려할 필요가 있다.

■ 일반화 선형모형에서는 약간 잠재된 연속형 종속변수를 j개의 순서화된 범주를 가지는 명백한 순서형 종속변수로 변환한다. 각 범주들을 구분하는 연속형 분포의 절단값들은 절편항 θ_j에 의해 추정된다. 어떤 경우에는 이러한 분포를 구분하는 이론적 근거가 있다. 그러나 심지어 잠재된 변수에 관한 이론적 개념이 없는 경우에도, 이 모형은 여전히 잘 맞으며 유효한 결과를 주는 경우가 있다. 또한 순서회귀모형에서 경계점(절단값)은 모형의 한 부분

으로 추정되며, 사전에 미리 명시할 필요가 없다.

■ 모형 내의 절편항 θ_j(경계점 또는 상수)는 오직 어떤 한 범주에 속할 확률을 예측하는 데 사용된다. 독립변수들의 값은 모형의 이 부분에 아무런 영향을 주지 않는다.

■ 모형의 예측에 해당하는 부분인 $(\beta_1 x_1 + \cdots + \beta_k x_k)$는 독립변수에만 의존하며 종속변수의 범주와는 무관하다.

■ 모형은 누적확률값이 아니라 이 값들의 함수값을 예측한다. 이런 함수를 연결함수라 하며, 모형을 설정할 때 결과를 최적화하기 위해 적절한 연결함수를 선택한다.

순서 로지스틱 회귀모형은 다음과 같이 세 개의 요소로 구성되어 있다.

■ 위치요소(location component)

식 (2.4)에서 회귀계수들과 독립변수들을 포함한 부분 $(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$을 모형의 위치요소라 한다. 이 요소에서는 한 개체가 어느 한 범주에 속할 확률을 예측하기 위한 독립변수를 사용한다.

■ 척도요소(scale component)

척도요소는 독립변수들의 변동이 서로 다른 경우에 선택할 수 있는 옵션이다. 예를 들어, 남성이 여성보다 은행계좌상태의 변동이 크다면 척도요소를 선택하여 모형을 개선할 수 있다. 척도요소가 있는 모형의 형태는

$$\text{link}(\gamma_j) = \frac{\theta_j - [\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k]}{\exp(\tau_1 z_1 + \tau_2 z_2 + \cdots + \tau_m z_m)}$$

이다. 여기에서, $\tau_1, \cdots, \tau_m$은 척도요소에 해당하는 계수들이며, $z_1, \cdots, z_m$은 척도요소에 대한 m개의 독립변수들이다. 이때 $z_1, \cdots, z_m$은 동일한 독립변수들의 집합에서 선택한다. 변동의 차이를 조정하기 위해 척도요소로 변수 성별뿐만 아니라 또 다른 독립변수들을 포함할 수 있다.

■ 연결함수

연결함수는 모형을 적절히 추정하기 위하여 누적확률을 변환시키는 것이다. 다음은 실제 상황에서 많이 사용하는 연결함수이다.

[표 1] 연결함수의 유형

함수	형태	용용
로짓	$\log\left(\frac{\gamma}{1-\gamma}\right)$	골고루 분포된 범주
보완 log-log	$\log(-\log(1-\gamma))$	높은 범수의 개체수가 더 많을 것 같은 경우
음의 log-log	$-\log(-\log(\gamma))$	낮은 범주의 개체수가 더 많을 것 같은 경우
프로빗	$\Phi^{-1}(\gamma)$	명백히 정규분포를 하는 잠재변수 분석시
cauchit (역 Cauchy)	$\tan(\pi(\gamma-0.5))$	극단값이 많은 종속변수

III. 사례연구 : 소비자 신용자료

본 장에서는 소비자 신용자료(Blake 외 2인, 1998)를 사용하여 2장에서 언급한 순서 로지스틱 회귀모형을 설정 및 평가하며, 더 나아가 다른 형태의 신용자료에 그 쓰임도를 적용하게끔 하도록 한다. 종속변수는 순서형인 5개 수준(과거 부채없음, 현재 부채없음, 현재 부채상환, 부채상환 만기경과, 위험계좌)으로 된 은행신용상태(account status)이다. 잠재적인 독립변수로 연령, 신용 대출횟수, 주택유형, 은행계좌 상태 등을 포함한 지원자의 다양한 재정적, 개인적인 특성으로 구성한다. 참고로 순서 로지스틱 회귀모형을 자료분석에 직접 적용하기 위해서 통계패키지 SPSS 프로시지의 PLUM을 사용하였다.

1. 모형 설정

순서 로지스틱 회귀모형을 설정하기 위해서 우선 순서형 결과변수를 식별한다. 그리고 잠재적인 독립변수 중 어떤 독립변수를 모형의 위치요소에 사용해야 할 지를 결정한다. 다음은 척도요소를 사용할 지의 여부를 정하고, 만일 사용한다면 어떤 독립변수에 사용할 지 결정한다. 마지막으로 연구의 초점과 자료의 구조상 어떤 연결함수를 사용하여야 하는 지를 선택한다.

2. 종속변수 식별

여기에서 5개의 범주(과거 부채없음, 현재 부채없음, 현재 부채상환, 부채상환 만기경과, 위험계좌)를 갖는 은행신용상태가 순서형 종속변수가 된다.

3. 위치모형에 대한 독립변수 선택

가정한 모형의 위치요소에 대한 독립변수를 선택하는 과정은 선형회귀모형에서 독립변수를 선택하는 과정과 흡사하다. 독립변수를 선택하려면 이론적인 면과 실제적인 면을 동시에 고려해야 한다. 이상적으로는 모든 중요한 독립변수만을 모형에 포함하여야 하지만, 실제로 모형을 설정할 때까지 어떤 독립변수가 중요한 지 정확히 알 수가 없다. 만일 어떤 독립변수가 모형에 필요없다고 판단되면 이들을 제거하고 모형을 다시 추정한다.

이 사례에서는 지난 경험과 탐색적 자료분석에 의해 5개의 독립변수(연령, 대출기간, 신용대출횟수, 다른 월부금, 주택유형)를 선택하였다. 처음에는 5개의 독립변수를 모두 포함하여 각 변수들의 중요성을 평가한다. 신용대출횟수, 다른 월부금, 주택유형은 범주형 독립변수로 모형에 요인으로 입력한다. 연령과 대출기간은 연속형 독립변수로 모형에 공변량으로 입력한다.

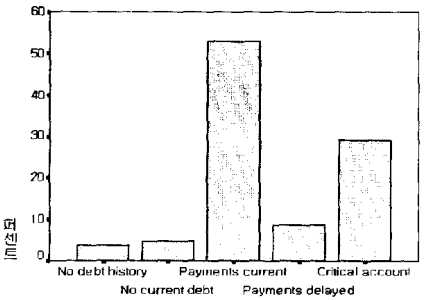
4. 척도요소

척도요소에는 두 단계 설정사항이 있다. 첫째는 모형에 척도요소를 포함할 지 여부를 결정하는 것이다. 대부분의 경우 척도요소는 필요하지 않고 위치요소만 포함한 모형이 자료를 잘 요약해 준다. 분석을 단순화하기 위해서는 위치요소만 포함한 모형부터 시작하는 것이 가장 좋다. 위치요소만을 포함한 모형이 주어진 자료에 대해 부적절한 경우 척도요소를 모형에 추가시킨다. 이러한 사실에 착안하여 위치요소만을 포함한 모형으로 시작하고 모형을 추정한 뒤 척도요소가 필요한 지의 여부를 점검한다.

5. 연결함수 선택

연결함수를 선택하기 위해서는 종속변수의 분포를 조사하는 것이 도움이 된다. [그림 1]은 신용상태의 범주에 대한 분포를 나타낸 것이다.

[그림 1] 신용상태 범주에 대한 분포



[그림 1]에서 개체의 대부분이 범주 3(현재 부채상환)과 범주 5(위험계좌)와 같이 높은 범주에 있다. 가장 중요한 특징들이 범주 3, 4, 5에 있기 때문에 높은 범주들에서 대부분의 조치를 취하여야 할 것이다. 이런 이유로 함수가 종속변수의 높은 범주들에 초점을 두어야 하므로 보완 log-log 연결함수로 시작한다. 범주 5(위험계좌)에 해당하는 개체의 수가 많은 편이므로 만일 보완 log-log 연결함수를 사용하여 만족스러운 결과를 얻지 못하면 대안으로 역 Cauchy 분포를 사용할 수 있다.

6. 모형의 예측값

각각의 독립변수를 살펴보기 전에 가정한 모형이 종속변수를 잘 설명하고 있는 지를 여부를 조사하기 위해서 [표 2]를 살펴보자. 이 표에서 절편항 만 있는 모형과 절편과 독립변수를 포함한 최종모형에 대한 -2(로그 가능도비)가 주어져 있으며, 로그 가능도값의 차이는 카이제곱 분포를 한다(McCullagh & Nelder, 1989).

[표 2] 모형적합도 정보

모형	-2 log 우도	MFI 카이제곱	자유도	MFI 유의확률
절편 만	2249.888			
최종	1896.552	353.336	9	.000

링크 함수: 보완 로그-로그.

카이제곱 통계량이 통계적으로 유의한 값을 가지면, 절편만을 포함한 모형에 대해 최종모형이 유의하게 향상되었음을 의미한다. 즉, 종속변수의 범주에 대하여 독립변수없이 주변확률들에 근거한 것보다 이 모형을 통해 더 좋은 예측값을 얻을 수 있음을 의미한다.

7. 카이제곱에 근거한 적합통계량

다음에 주어진 [표 3]에 나타난 두 개의 통계량들을 통하여 적합된 모형이 자료를 얼마나 잘 설명하는 지를 알 수 있다. p값이 유의하게 크면, 가정된 모형이 자료를 잘 설명함으로 좋은 모형이라고 결론지을 수 있다.

[표 3] 적합도

	GOF 카이제곱	자유도	GOF 유의확률
Pearson	4688.724	3131	.000
편차	1796.915	3131	1.000

링크 함수: 보완 로그-로그.

이 통계량들은 수준의 수가 적은 독립변수를 가진 모형에 매우 유용할 수 있다. 그러나 이 통계량들은 빈 셀에 민감하다. 연속형 공변량이 있는 모형을 추정할 때 빈 셀이 많이 존재한다. 그러므로 이런 모형에서 구한 두 개의 통계량들은 모두 신뢰해서는 안될 것이다. 빈 셀로 인해 두 통계량들이 실제로 카이제곱 분포를 한다고 확신할 수 없으며 따라서 유의확률은 정확하지 않게 된다.

8. 유사 R^2 통계량

모형의 전반적인 적합도를 평가할 수 있는 또 다른 도구는 유사 R^2 통계량이다. 이 통계량은 선형회귀모형에서 결정계수와 같은 역할을 한다. 즉, 독립변수(들)와 관련된 종속변수에서의 분산의 비를 요약한 것이다. 순서 로지스틱 회귀모형에서 이 통계량은 잔차가 아닌 가능도비에 근거한다. [표 4]는 초기모형에 대한 결과이다. 세 개의 다른 형태의 결정계수를 알아보자.

[표 4] 유사 R^2 값

Cox와 Snell	.298
Nagelkerke	.328
McFadden	.149

링크 함수: 보완 로그-로그.

Cox & Snell의 R^2(Cox와 Snell, 1989)은 최대가능도 추정법을 사용할 때 기존의 R^2을 일반화한 것이다. 그러나 종속변수가 범주형일 때 이론적으로 Cox & Snell의 R^2은 "1.0"보다 작게 된다. 이런 이유로 Nagelkerke(1991)가 제안한 통계량은 0에서 1의 값을 갖도록 수정된 R^2을 제안하였다. McFadden의 R^2(McFadden, 1974)은 절편항만 있는 모형과 절편항과 모든 독립변수를 포함한 모형의 로그 가능도에 근거한 또 다른 형태의 R^2을 제안하였다.

여기에서 계산된 유사 R^2 통계량의 값은 어느 정도 크지만 만족스럽지는 않다. 즉, 예측을 더 좋게 하기 위해 모형을 변환할 필요가 있다. 모형을 평가하기 위한 다음 절차는 모형에 의해 생성된 예측값을 조사하는 것이다. 모형은 누적확률을 예측하는데 기반을 두고 있다. 그러나 가장 큰 관심은 가성한 모형이 독립변수의 값에 근거하여 얼마나 빈번하게 범주들을 정확하게 예측하는 지의 여부이다. 모형이 얼마나 좋은 지 알기 위해서는 예측범주를 실제 범주와 교차 분류하여 분류표(classification table)를 만드는 것이다. [표 5]는 이 사례에 해당하는 분류표이다.

[표 5] 초기모형에 대한 분류표

			예측된 응답범주		전체
			3	5	
Account Status	No debt history	빈도	14	26	40
		Account Status의 %	35.0%	65.0%	100.0%
	No current debt	빈도	41	8	49
		Account Status의 %	83.7%	16.3%	100.0%
	Payments current	빈도	480	50	530
		Account Status의 %	90.6%	9.4%	100.0%
	Payments delayed	빈도	31	57	88
		Account Status의 %	35.2%	64.8%	100.0%
	Critical account	빈도	73	220	293
		Account Status의 %	24.9%	75.1%	100.0%
전체		빈도	639	361	1000
		Account Status의 %	63.9%	36.1%	100.0%

가정한 모형은 적어도 가장 빈도의 수가 많은 범주 3과 범주 5에서 종속변수의 범주를 예측하는데 상당한 역할은 한 것 같다. 모형은 범주 3의 개체들의 90.6%를, 범주 5의 개체들의 75.1%를 정확히 분류하였다. 추가로 범주 2의 개체들은 범주 5 보다는 범주 3으로 분류될 가능성이 매우 높다.

한편 범주 1(신용거래 없음)의 개체들은 어느 정도 잘 예측되지 않았다. 이는 순서형 결과변수의 척도를 정하는데 문제가 있음을 의미한다. 이를 더 진행시키지 않겠지만, 실제 자료분석 상황에서는 이에 중점을 두어 범주들을 재 정렬, 병합, 제거함으로써 순서척도를 개선하여야 할 것이다.

9. 문제식별

1) 연결함수의 선택

명확한 이론으로 자료에 근거한 연결함수를 선택할 수는 없다. 초기 모형이 적절하지 못한 경우에는 다른 연결함수를 설정하여 더 좋은 모형을 만들 수 있을 지 여부를 타진해 볼 수 있겠다. 몇 개의 연결함수가 많은 경우에 있어 거의 비슷한 결과를 보여 주지만(특히 로짓, 보완 log-log, 음 log-log), 어떤 연결함수를 사용하느냐에 따라 모형을 더 좋게 혹은 나쁘게 만들 수 있는 상황들이 발생한다.

이 사례에서는 적어도 두 개의 적절한 연결함수(보완 log-log, 역 Cauchy)를 고려할 것이다. 모형에는 보완 log-log 연결함수가 잘 맞을 지라도, 역 Cauchy 연결함수를 사용하여 모형을 더 좋게 만들 수도 있다.

2) 평행선 검정

위치요소만을 포함한 모형에 대해 평행선 검정을 통해 모든 범주들에 대한 모수들이 동일하다는 가설이 사실인 지를 평가할 수 있다. 이 검정은 모든 범주에 대한 계수들을 포함한 추정모형과 각 범주에 대해 개별적인 계수를 갖는 모형을 비교한다. 즉,

$$\text{link}(\gamma_j) = \theta_j - [\beta_{1j}x_1 + \beta_{2j}x_2 + \cdots + \beta_{kj}x_k]$$

와

$$\text{link}(\gamma_j) = \theta_j - [\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k]$$

를 비교한다. [표 6]은 이 사례에 대한 결과이다. 각 범주에 대해 개별적인 모수를 가진 일반화된 모형이 그렇지 않은 모형에 비해 통계적으로 유의하게 나아졌음을 알 수 있다. 올바르지 않은 연결함수 또는 잘못된 모형을 사용하였으므로 이런 상황이 발생하였을 것이다.

[표 6] 평행선 검정

모형	-2log 우도	TPL 카이제곱	자유도	TPL 유의확률
영가설	1896.552			
일반	1588.614[a]	307.938[b]	27	.000

영가설 상태는 위치 모수가(기울기 계수) 대응 범주에 있어 동일함을 나타냅니다.
a. 최대 반분 수 이후에 로그-우도값을 더 이상 증가시킬 수 없습니다.
b. 카이제곱 통계량은 일반 모형의 마지막 반복계산 로그-우도값을 기준으로 계산합니다. 검정의 유효성은 확실하지 않습니다.
c. 링크 함수: 보완 로그-로그.

3) 모형 내의 독립변수들

[표 7] 모수 추정값

		B 추정값	표준 오차	Wald	자유도	PAR 유의확률	95% 신뢰구간	
							하한	상한
한계치	[A3 = 0]	-3.549	.667	28.323	1	.000	-4.856	-2.242
	[A3 = 1]	-2.720	.656	17.167	1	.000	-4.006	-1.433
	[A3 = 2]	-.137	.649	.044	1	.833	-1.408	1.135
	[A3 = 3]	.199	.649	.094	1	.759	-1.072	1.471
위치	A13	1.500E-02	.004	15.128	1	.000	7.441E-03	2.256E-02
	A2	-2.08E-03	.003	.379	1	.538	-8.69E-03	4.534E-03
	[A16=1]	-1.134	.594	3.645	1	.056	-2.298	3.019E-02
	[A16=2]	.367	.598	.376	1	.540	-.805	1.538
	[A16=3]	.981	.711	1.902	1	.168	-.413	2.374
	[A16=4]	0[a]	.	.	0	.	.	.
	[A14=1]	-.397	.118	11.389	1	.001	-.627	-.166
	[A14=2]	-.469	.193	5.913	1	.015	-.848	-9.11E-02
	[A14=3]	0[a]	.	.	0	.	.	.
	[A15=1]	-8.25E-02	.165	.249	1	.617	-.406	.241
	[A15=2]	.132	.139	.897	1	.344	-.141	.404
	[A15=3]	0[a]	.	.	0	.	.	.

링크 함수: 보완 로그-로그.
a. 현재 모수는 중복되므로 0으로 설정됩니다.

이제 모형 내의 각 독립변수들을 고려해 보자. [표 7]은 모수 추정값들이다. 흔히 있는 것처럼 이 사례에서 절편항들은 이론적인 면으로 볼 때 중요하지 않다. 가장 관심있는 모수들은 독립변수들과 종속변수 범주의 누적확률을 모형화하는 위치모수들이다.

모수 추정값들과 이에 관련된 통계적 검정을 통하여 어느 독립변수를 모형으로부터 제거할 지를 결정한다. 예를 들어, 변수 대출기간은 이 모형에서 중요하지 않아 보인다. 일단 적절한 연결함수를 설정하였다면 공변량의 유의성을 재평가하여 가능한한 모수를 절약한 모형을 만든다. 변수 주택유형은 모형에 대해 그 유의성 여부가 확실하지 않으므로, 모형을 바꾼 후에 이 변수를 재평가해야 한다.

10. 모형 수정

역 Cauchy 연결함수를 사용한 새로운 모형을 추정하여 모형의 예측력을 증가시켰는지를 알아보도록 하자. [표 8]의 모형 적합도 통계량을 통해 이 모형이 단순히 추측한 것보다 좋다는 것을 알 수 있다. 그러나 절편항 만을 포함한 모형과 절편항과 모든 독립변수를 포함한 모형을 비교한 카이제곱 통계량을 보면 역 Cauchy 연결함수를 사용한 경우 ($\chi^2 = 459.860$)가 보완 log-log 연결함수를 사용한 경우 ($\chi^2 = 353.336$)보다 매우 큼을 알 수 있다. [표 9]는 유사 R^2 통계량들의 값들이다.

[표 8] 모형 적합도 정보

모형	-2 log 우도	MFI 카이제곱	자유도	MFI 유의확률
절편 만	2249.888			
최종	1790.028	459.860	9	.000

링크 함수: Cauchit.

[표 9] 유사 R^2 값

Cox와 Snell	.369
Nagelkerke	.407
McFadden	.194

링크 함수: Cauchit.

이를 통해 연결함수를 역 Cauchy로 바꾸어 모형이 종속변수 내에 존재하는 패턴의 설명력을 향상시켰다는 것을 알 수 있다.

[표 10]은 분류표이다. 이 모형이 보완 log-log 연결함수를 사용한 모형에 비해 낮은 범주(1,2,3)는 보다 잘 예측하지만 높은 범주를 잘 예측하지 못한다. 신용을 점수화하는 가장 중요한 목표는 위험계좌(범주 5)가 될 것 같은 계좌를 정확히 식별하는데 있기 때문에, 설사 수정된 모형에 대한 적합통계량이 만족스럽다 하더라도 보완 log-log 연결함수를 사용한 원래의 모형을 선택할 것이다.

[표 10] 분류표

			예측된 응답범주		전체
			3	5	
Account Status	No debt history	빈도	15	25	40
		Account Status의 %	37.5%	62.5%	100.0%
	No current debt	빈도	43	6	49
		Account Status의 %	87.8%	12.2%	100.0%
	Payments current	빈도	482	48	530
		Account Status의 %	90.9%	9.1%	100.0%
	Payments delayed	빈도	36	52	88
		Account Status의 %	40.9%	59.1%	100.0%
	Critical account	빈도	80	213	293
		Account Status의 %	27.3%	72.7%	100.0%
전체		빈도	656	344	1000
		Account Status의 %	65.6%	34.4%	100.0%

11. 모형해석

모형이 자료에 적절히 적합되면 [표 7]에서 보는 바와 같이 모수의 추정값에 근거하여 해석할 수 있다. 통계적 검정에 따르면 변수 대출기간은 모형에서 제거되고 변수 연령은 유의한 변수임을 알 수 있다. 변수 신용대출횟수의 수준들은 모두 유의하지 않지만 유의확률이 5%에 가까운 것은 두 개가 있다. 일반적으로 각 범주들의 조그마한 효과들이 모여 모형에 유용한 정보를 주기 때문에 이러한 변수는 남겨둘 가치가 있다. 변수 다른 월부금 역시 중요한 독립변수인 듯 하다. 반면에 변수 주택유형은 그 스스로 의미가 있기 때문에 통계적으로 유의하지 않더라도 모형에 포함시켜야 할 것이다.

모형에서 계수들을 직접적으로 해석하는 것은 연결함수의 성질 때문에 쉽지는 않지만, 계수들의 부호를 통해 모형 내의 독립변수의 효과를 인식할 수 있다. 추정된 계수의 부호는 효과의 방향을 나타낸다. 변수 연령의 계수가 양의 부호를 가지므로 독립변수들과 종속변수 간에 양의 관계를 나타낸다. 이 경우 변수 연령의 값이 증가할수록 종속변수 은행계좌상태의 높은 범주들 중 하나에 속할 확률이 증가한다. 변수 신용대출횟수의 첫 번째 범주와 같

이 부호가 음이면 역관계를 나타낸다. 예를 들어, 이 모형에서는 신용대출횟수가 하나인 개체들은 종속변수 은행계좌상태의 낮은 범주에 속할 가능성이 크게 된다.

12. 모형의 예측

1) 예측확률

모형은 종속변수의 한 범주가 아닌 누적된 확률을 예측하므로 예측범주들을 예측하기 위해서는 두 단계의 과정이 필요하다. 첫째 각 개체에 대해 각 범주에 속할 확률을 각각 추정한다. 둘째, 이 확률들을 이용하여 각 개체에 대해 종속변수의 범주 중 가장 확률이 높은 범주를 선택한다.

가정된 모형에서 각 개체에 대한 독립변수의 값을 사용하여 연결함수의 역변환을 통해 확률을 추정한다. 따라서 각 개체에 대한 독립변수들의 값을 이용하여 각 그룹에 대한 누적확률을 얻게 된다. 각 범주에 대한 확률은 순서대로 나열된 범주에 대한 누적확률의 차이를 이용하여 구한다. 즉, 처음 범주에 대한 확률은 첫 번째 누적확률이다. 두 번째 범주에 속할 확률은 두 번째 누적확률에서 첫 번째 누적확률을 빼면 된다. 나머지 범주에 속할 확률도 같은 방법으로 구하면 된다.

2) 확률값을 이용한 범주의 예측

각 개체에 대한 종속변수 범주의 예측은 단순히 그 개체의 독립변수의 값이 주어졌을 때 가장 큰 확률을 갖는 범주가 된다. 예를 들면, 변수 대출기간의 값이 48개월이고, 연령이 22세, 한 개의 신용대출횟수, 다른 월부금이 없고, 변수 주택을 소유한 개체를 생각하여 보자. 이 값들을 추정한 식에 대입하면 각각 -2.78, -1.95, 0.63, 0.97의 예측된 값을 얻게 된다. 보완 log-log 연결함수의 역변환을 통하여 각각 0.06, 0.13, 0.85, 0.93인 누적확률을 얻게된다. (물론 이때 마지막 범주에 대한 누적확률은 1.0이다.) 따라서 다음과 같은 개별적인 범주의 확률을 구할 수 있다.

범주 1 : 0.06

범주 2 : 0.13 − 0.06 = 0.07

범주 3 : 0.85 − 0.13 = 0.72

범주 4 : 0.93 − 0.85 = 0.08

범주 5 : 1.00 − 0.93 = 0.07

분명히, 범주 3(현재 부채상환)은 이 개체의 경우 0.72의 예측확률을 갖게 된다. 따라서 이 개체는 계속해서 상환금을 갚고 은행계좌에 문제가 발생하지 않을 것이라고 예측할 수 있다.

Ⅳ. 결론

본 논문에서는 신용거래점수모형으로 사용할 수 있는 순서 로지스틱 회귀모형을 소개하였으며, 신용거래에 대한 의사결정을 하는데 이 모형이 어떻게 적용되는 지를 소비자 신용자료를 통해 그 응용도를 살펴보았다. 순서 로지스틱 회귀분석은 실제자료에 근거한 통계적 방법이기 때문에 주관적인 판단방법에 비해 더 신뢰할만하다. 특히 이 사례연구에서 사용된 독립변수들을 이에 관련된 다른 상황에서 절대시하여 적용시킬 수 없다. 신용위험을 적절하게 설명할 수 있는 설명변수를 선택하며, 선택된 설명변수 각각에 신용거래점수 기여도 정도에 따라 가중치를 주어야 할 것이다.

참 고 문 헌

[1] Blake, C., Keogh, E. and Merz, C. J., UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLR Repository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[2] Cox, D. R. and Snell, E. J., Analysis of Binary Data, 2nd ed., London: Chapman & Hall, 1989.

[3] Hosmer, D. W. and Lemeshow, S., Applied Logistic Regression, New York: John Wiley and Sons, 2000.

[4] McCullagh, P. and Nelder, J. A., Generalized linear models, 2nd ed., New York: Chapman & Hall, 1989.

[5] McFadden, D., Conditional logit analysis of qualitative choice behavior. In Zarembka, P. (ed.), Frontiers in Econometrics. New York: Academic Press, 1974.

[6] Nagelkerke, N. J. D., "A note on general definition of the coefficient of determination," Biometrika, Vol.78, 1991, pp.691-692.

ABSTRACT

Study on Credit Scoring Model
- Ordinal Logistic Regression Model -

Jeong, Dong-Bin*

A credit score is a number that reflects credit risk level. typically with a higher number indicating lower risk. In this study, the ordinal logistic regression model is introduced, which is one of the most successful credit scoring models. In addition, real consumer credit data are applied to the ordinal logistic regression model as a case study. By analyzing and evaluating this model, we can make decision on a variety of credit things such as credit ceiling, credit loan and so on and furthermore partly understand the economic phenomenon and mechanism.

Keyword *: credit scoring, link function, maximum likelihood estimator*

* Associate Professor, Department of Information Statistics, Kangnung National University