Journal of Korean
Data & Information Science Society
2008, Vol. 19, No. 3, pp. 735 ~ 749

Analysis of Multicategory Responses with Logit Model on Earlyold Age Pension[1)]

Mijung Kim[2)]

Abstract

This article suggests application of logit model for analysis of multicategory responses. Referring to the reference category, characteristic of each category is obtained from analysis of polytomous logit model. With National Pension data it is illustrated that application of logit model helps it possible to find significant factors which may not be found only with polytomous logit model. Application of the logit model is done by reducing the number of categories. Categories are grouped into the former and the latter group according to reference category. Extra finding of significant factor was possible from logistic regression analysis for the two groups after removing the reference category. It is expected that this application would be helpful for finding information and characteristics on ordered multicategory responses where the proportional odds model does not fit.

Keywords: Earlyold Age Pension; Logit model; National Pension; Ordered multicategory responses; Polytomous logit model.

1. 머리말

범주형 자료는 변수가 가질 수 있는 값이 명목형(nominal) 또는 순위형 척도인 자료를 말하며, 이러한 자료 분석은 각 범주에 속하는 상대적인 빈도, 즉 비율을 사용한다. 로짓모형은 종속변수가 범주형인 경우 회귀모형을 적용하여 각 범주에 할당될 확률을 예측함으로써 관측치가 각 범주에 속할 확률을 해석하거나 각 범주에 대한 threshold를 기준으로 관측치를 특정범주로 분류할 때 그리고 범주형 반응변수에 대한 설명변수의 회귀계수에 대한 통계적 유의성을 해석하기 위한 목적으로 적용한다. 여기서 로짓모형은 다범주 자료를 적합하기 위한 모형으로 비례오즈모형과 다항로짓모

1) 연구 자료는 국민연금관리공단의 공동연구 자료임.
2) 서울시 서대문구 신촌동 134번지 연세대학교 수리과학 연구소 연구교수
E-mail: mjkim@yonsei.ac.kr

형, 그리고 로지스틱회귀모형을 포함한다.

국민연금의 노령연금가운데 하나인 조기노령연금은 가입기간이 10년 이상이고 55세 이상인 사람이 소득이 없는 경우 본인이 신청하면 60세 이전이라도 지급받을 수 있는 연금이다. 따라서 조기노령연금을 수급할 수 있는 연령은 55세부터 59세까지로 다섯 개의 순위형 척도로 파악될 수 있다.

본 연구는 조기노령연금 수급자를 대상으로 한 조기노령연금의 특성 분석으로 조기노령연금의 수급개시연령에 대한 설명요인의 특성을 파악하기 위하여 로짓 모형을 응용하였다. 즉, 조기노령연금의 수급 가능 연령은 55세부터 59세까지로, 소득수준, 소득활동여부, 가입종별, 성별을 설명변수로 하였을 때 수급개시 연령의 선택에 영향을 주는 요인의 특성을 파악함으로써 설명요인의 값의 변화에 따라 이른 수급 혹은 늦은 수급으로 갈 가능성이 어느 만큼 높아지는가를 파악할 뿐 만 아니라 수급연령의 선택에 있어서 이른 수급 또는 늦은 수급의 일정한 방향으로 영향을 주는 설명요인의 특성 파악 - 본문에는 '수급연령의 증감에 대한 설명요인의 '일정한' 특성파악'으로 서술함 -을 위하여 수급개시연령을 반응변수로, 4개의 요인들을 설명변수로 하여 다항로짓 모형을 적용하되 필요에 따라 범주를 그룹화하거나 제거하는 방법을 사용하였다.

분석 목적가운데 하나인 수급연령의 증감에 따른 설명요인의 '일정한' 특성파악은 다섯 범주에 대한 다항로짓 모형으로는 파악이 어려웠기 때문에 범주를 '이른 수급'과 '늦은 수급'의 경우로 그룹화 하였고 분석절차에 있어서, 오분류율을 낮추기 위하여 정분류율이 높은 범주를 제외한 특정범주를 대상으로 판별분석을 시도하는 데이터마이닝의 기법을 응용하여, 기준범주를 제외하고 그룹화한 두 개의 그룹에 대한 로짓분석을 시도하였다.

분석 자료로 주어진 국민연금관리공단의 조기노령연금 수급자 자료는 조기노령연금의 수급연령에 대한 비례오즈모형 가정 사항을 만족하지 않아 비례오즈모형 적용이 불가능하였고 따라서 다항로짓 모형을 적용하는 것이 일반적이나, 범주가 많아 다항로짓 분석만으로 다섯 범주 값의 증감에 따른 설명요인의 일정한 추이를 살펴보는 데 어려움이 있었다. 따라서 아래에 소개한 바와 같이, 필요에 따라 범주를 그룹화하거나 제거하여 로짓모형에 적합하는 방법을 사용하였다.

먼저 다섯 범주의 순위형 변수를 종속변수로 하는 다항 로짓 모형을 적용하였다. 다항 로짓은 종속변수의 여러 범주 중에서 한 범주를 기준범주로 정하고, 다른 범주를 기준범주와 비교하는 방법을 사용한다. 즉, 종속변수의 결과가 기준범주에 비하여 다른 범주에 속할 확률을 계산하는 것이다 (Maddala, 1983). 55-59까지의 다섯 개의 종속변수의 범주 가운데 대표값인 수급개시연령 57을 기준범주로 두고 타 범주를 비교하는 방법을 적용하였다. 이를 통하여 각 수급연령의 특성은 파악되었으나 이른 수급이거나 늦은 수급에 따른 설명요인의 일정한 특성파악이 어려웠으므로 55세와 56세 범주를 기준범주의 연령보다 비교적 일찍 수급하는 그룹 (former 그룹)으로 58세와 59세 범주를 기준범주의 연령보다 비교적 늦게 수급하는 그룹 (latter 그룹)으로 나누어 57을 기준 범주로 한 두 그룹의 특징을 살펴보았다. 이로부터, 57세를 기준으로 한 former 그룹과 latter 그룹 간에는 4개의 설명요인 모두 유의한 요인이기는 하였으나 성별을 제외한 다른 요인들은 설명요인의 값의 변화가 수급연령의 선택에 있어서 증가 또는 감소의 변화를 보이지 않음을 확인하였는데, 이는 기준범주의 특성이 타 범주들과 구별되는 특성을 보임으로 인하여 발생한 결과임을 확인하였다. 따라서 수급연령의 선택에 있어서 이른 수급 혹은 늦은 수급의 일정한 방향으로 영향을 주되

통계적으로 유의한 설명요인을 찾기 위하여 기준범주를 제외한 두 그룹에 대한 로지스틱분석을 추가적으로 시행하였다.

기준범주를 제거함으로써, 이른 수급과 늦은 수급의 선택에 있어서 성별뿐 아니라 가입종별도 수급연령의 증감에 대하여 일정한 방향을 보이는, 통계적으로 유의한 요인임을 추가적으로 확인하였는데, 이는 기준범주의 특성이 타 범주들의 특성과 다름을 확인한 후 이를 제외한 두 그룹만을 대상으로 로짓분석을 시행함으로써 얻을 수 있는 결과였다.

또한, 앞의 절차를 따라 확인한 가입종별의 특성에 대한 확인과정으로 다섯 개의 개별 수급나이에 대한 로지스틱분석을 수행함으로써 각각의 수급연령에 대한 설명요인의 특성을 파악하였다. 이러한 로지스틱 회귀분석으로부터, 가입종별은 모든 수급연령에 대하여 공통적으로 유의한 요인임을 확인하였으며 임의 가입에 비해 사업장가입이거나 지역가입일 경우 이른 수급일 가능성이 높음을 보여, 본 연구에서 제안하는 절차를 따라 기준 범주를 제외한 두 그룹에 대하여 시행한 로지스틱회귀분석과 같은 결과를 보임을 확인하였다. 즉, 조기노령연금 수급자가 그들의 가입종별 형태에 따라 이른 수급 혹은 늦은 수급의 일정한 방향의 선택을 하고 있으며 그러한 선택에는 '통계적으로' 유의한 차이가 있음을 확인할 수 있었는데 이는 본 연구가 제안하는 절차에 따른 분석으로부터 얻을 수 있었던 결과이다.

분석 결과가 보여주는 odds ratio는 설명요인의 값의 변화에 따라 이른 수급 혹은 늦은 수급의 선택으로 갈 가능성이 어느 만큼 높아지는지를 설명하므로, 조기노령연금의 재정 관리에 있어서 이른 수급과 늦은 수급 간 차이를 반영하기 위한 값으로 제시될 수 있는데, 예컨대 조기노령연금의 관리에 있어서 설명요인의 값의 변화에 따라 이른수급을 선택할 확률과 늦은 수급을 선택할 확률의 비율 - 분석결과로부터 얻은 odds ratio값 -을 이른 수급과 늦은 수급 간 차이를 반영하기 위한 값으로 적용할 수 있다. 본 연구의 결과가 제시하는 모형은 조기노령연금의 대상자 관리에 있어서 이른 수급과 늦은 수급의 대상자를 예측, 분류하고 조기노령연금 재정을 수급 대상자의 특성에 따라 관리하는데 도움을 줄 수 있을 것이다.

2. 실증분석을 위한 설계

2.1 분석프로세스

국민연금관리공단에서는 국민연금을 효율적으로 관리하고 모니터링하기 위해 주기적으로 가입종별, 성별, 소득활동여부, 그리고 소득등급에 대한 변화추이를 관찰하고 있으며, 각 연금에 따른 기금운용과 수급자들의 특성들을 파악하고 있다 (국민연금관리공단, 2007).

본 연구에서는 1988년 1월부터 2007년 5월까지의 국민연금가입자 가운데 조기노령연금 수급자를 대상으로 수급개시 연령에 따른 조기노령연금 수급자들의 특징들을 파악하기 위하여 55세부터 59세까지 선택 가능한 조기노령연금 수급 연령을 종속변수로 하여 소득수준, 소득활동여부, 가입종별, 성별을 설명변수로 하였을 때 수급개시 연령의 선택에 영향을 주는 요인의 특성을 파악하여 설명요인의 값의 변화에 따라 이른 수급 혹은 늦은 수급으로 갈 가능성이 어느 만큼 높아지는가를 파악하고 또한 수급연

령의 증감에 대하여 통계적으로 유의한 설명요인을 파악하기 위하여 다항로짓 모형을 적용하였다.

우선 다섯 범주에 대한 다항로짓 분석을 실시하였고, 이를 통하여 57을 기준으로 각 범주에 속하는 특징들을 살펴보았으며, 55와 56을 former 그룹으로, 58과 59를 latter 그룹으로 구분하여 57을 기준으로 former 그룹과 latter 그룹에 따른 특징을 살펴봄으로써 조기노령연금의 수급개시 평균연령인 57세보다 일찍 수급하는 경우와 늦게 수급하는 경우의 특성을 파악하고자 하였다. 세 그룹에 대한 다항로짓 모형 외에, former 그룹과 latter 그룹의 특성을 파악하기 위하여 57세 범주를 제외한 두 그룹에 대한 로지스틱 회귀 모형을 적용하였다.

이로부터 기준범주인 57세 범주 대비 각각 네 개의 수급연령 범주가 가지고 있는 특성과, 57세 범주 대비 former 그룹과 latter 그룹의 특성뿐 만아니라, 기준 범주인 57세 범주의 특성, 그리고 former 그룹과 latter 그룹에 대해 유의한 차이를 보이는 요인들을 찾을 수 있었다.

2.2 데이터의 구성

국민연금관리공단에서 관리하고 있는 조기노령연금 수급자 115,044명 가운데 성별 파악이 불가능한 3명의 자료값을 제거한 115,041명의 데이터를 사용하여 수급개시연령에 따른 조기노령연금의 수급자들에 대한 특징을 파악하였다. 분석에 사용된 변수들은 국민연금관리공단에서 주기적으로 모니터링하고 있는 성별, 소득활동여부, 소득수준, 그리고 가입종별을 설명변수로 사용하였으며, 수급개시연령을 종속변수로 사용하여 로짓 모형을 세웠다. 즉, 조기노령연금의 수급이 55세에서 59세까지 가능한 경우, 소득수준, 소득활동여부, 가입종별, 성별을 설명변수로 하였을 때 수급개시 연령의 선택에 영향을 주는 요인의 특성을 파악함으로써 설명요인의 값의 변화에 따라 이른 수급 혹은 늦은 수급으로 갈 가능성이 어느 만큼 높아지는가를 파악할 뿐 만 아니라 수급 연령의 선택에 있어서 이른 수급 또는 늦은 수급의 일정한 방향으로 영향을 주는 설명요인의 특성 파악을 위하여 수급개시연령을 반응변수로, 4개의 요인들을 설명변수로 하여 다항로짓 모형을 적용하였다.

사용된 변수에 대한 정보는 [표 1]과 같다. 수급개시연령은 가입자가 처음으로 수급했을 때의 가입자의 나이이며 각 변수들은 수급받기 직전의 가입 상태를 사용하여 분석을 실시하였다.

데이터를 살펴보면 수급개시연령은 조기노령연금을 수급하기 시작한 연령으로 55세는 33.87%, 56세는 21.50%, 57세는 19.28%, 58세는 17.01%, 그리고 59세는 8.35%로 구성되어 있다. 또한 성별의 경우 여성이 28.69%, 남성이 71.31%이며, 소득활동 여부의 경우, 취득은 소득활동이 있는 국민연금 가입자의 경우로 17.21%, 납부예외는 현재 국민연금 가입자 범주에 속하지만 소득활동이 없어서 국민연금 납부가 일정기간 유예된 자로서 1.30%, 그리고 대기는 국민연금 가입이력을 가지고 있으나 현재 소득활동이 없는 관계로 현 가입자 범주에 해당되지 않는 자이면서 수급 발생이 일어나지 않은 자로 81.49%의 비율로 분포되어 있다. 또한 가입종별의 경우, 사업장가입은 근로소득이 있는 직장 가입자로 전체의 39.99%를 지역가입이 55.96%, 그리고 임의가입은 사업장가입과 지역가입을 제외한 모든 경우에 해당하는 자로 4.05%로 나타났다.

[표 1] 로짓 모형에 사용된 변수

변수명	변수설명	변수의 특성
Age	수급나이	55,56,57,58,59
Sex	성별	1=남자, 0=여성
Status	소득활동여부	0=취득, 1=납부예외 2=대기
Wage	소득수준	1-45등급
Class	가입종별	0=사업장 가입, 1=지역가입, 2=임의 가입

소득수준의 경우 소득이 있는 경우 표준 보수액을 기준으로 하여 1등급부터 45등급으로 구분되며, 1988년 1월 이후부터 95년 4월까지는 53등급으로 나누던 것을 95년 4월 이후에는 45등급으로 조정되어 분석에서는 95년 4월까지의 등급을 45로 보정하여 평균소득수준을 계산하였다. 보정할 때 사용한 방법은 53등급으로 나누었을 때의 53등급과 45등급으로 나누었을 때의 45등급이 동일한 등급임을 가정하였으며 직선보간(linear interpolation)을 하여 45등급으로 맞추었다. 사용된 변수에 대한 분포 및 기초 통계량 값은 [표 2]에 제시하였다.

[표 2] 사용된 변수에 대한 분포 및 기초 통계량

구분	변수	빈도	백분율(%)
수급나이	55	38961	33.87
	56	24729	21.50
	57	22179	19.28
	58	19568	17.01
	59	9604	8.35
성별	남성	82041	71.31
	여성	33000	28.69
소득활동여부	취득	19796	17.21
	납부예외	1495	1.30
	대기	93750	81.49
가입종별	사업장가입	46002	39.99
	지역가입	64377	55.96
	임의가입	4662	4.05

변수	자료수	평균값	표준편차	최소값	최대값
소득수준	115041	29.51	8.89	1.28	45

3. 모형설정

3.1 다섯 개 순위형 척도에 대한 다항 로짓 분석

셋 이상 범주의 순위형 반응변수의 경우 사용되는 로짓 모형은 비례오즈모형이다. 여기서 비례오즈란 각 범주에 속하는 odds들이 순위가 변함에 따라 동일하다는 뜻이다. 분석에서 사용한 종속변수는 다섯 개의 수급개시연령으로 순위형 반응변수이다. 비례오즈모형의 타당성 여부를 검정한 결과 비례오즈모형 가설에 대한 p-value가 <0.001로 비례오즈모형가설이 기각되었다. 따라서 비례오즈 모형을 적용할 수 없었고, 셋 이상의 범주를 갖는 명목형 반응변수에 대한 다항 로짓 모형을 적용하였다 (박용규, 2001). 이를 분석하기 위하여 SAS 9.1의 CATMOD 프로시져를 이용하여 분석을 실시하였다.

다항 로짓모형을 적용하면 기준 범주 대비 각 범주에 속하는 확률로 각 범주의 특성을 파악할 수 있다. 본 연구에서는 기준범주를 57세로 두어 수급연령이 56세, 55세로 기준범주보다 이른 수급이거나 58세, 59세로 기준범주보다 늦은 수급으로 갈수록 달라지는 수급자들의 특징을 파악하고자 하였다.

위의 변수들을 적용한 로짓 모형은 다음과 같다.

$$\log\left(\frac{\hat{\pi}_i}{\hat{\pi}_{57}}\right) = \beta_{i0} + \beta_{i1} \times Wage + \beta_{i2} \times J1 + \beta_{i3} \times J2 + \beta_{i4} \times S1 + \beta_{i5} \times S2 + \beta_{i6} \times Sex$$
$$i = 55, 56, 58, 59$$

여기서 $\hat{\pi}_i$는 수급연령이 i 에 속하는 확률의 추정치이며, J1은 가입종별이 사업장 가입일 경우 1 그렇지 않을 경우 0을 J2는 가입종별이 지역가입일 경우 1 그렇지 않을 경우 0을 나타낸다. S1은 소득활동여부가 취득일 경우 1 그렇지 않을 경우 0을 그리고 S2는 납부예외일 경우 1을 그렇지 않을 경우 0을 나타낸다.

[표 3]는 다항 로짓 모형에 대한 β의 추정치이며 이를 통하여 기준범주 57과 각 범주에 대한 특성들을 파악할 수 있다.

[표 3] 다항 로짓 모형에 대한 추정치

		범주 57과의 관계											
		55			56			58			59		
		estimate	standard error	p-value	estimate	standard error	p-value	estimate	standard error	p-value	estimate	standard error	p-value
Intercept		0.0988	0.0610	0.1052	−0.0163	0.0666	0.8071	−0.1427	0.0702	0.0420	−0.8548	0.0872	<.0001
Wage		0.0046	0.0012	0.0002	0.0004	0.0014	0.7454	0.0011	0.0014	0.4255	0.0049	0.0018	0.0061
J1	0	−0.0458	0.0444	0.3029	0.0928	0.0485	0.0557	−0.0222	0.0512	0.6642	−0.1871	0.0637	0.0033
J2	1	0.3367	0.0439	<.0001	0.0521	0.0481	0.2790	0.0251	0.0507	0.6213	0.0176	0.0628	0.7796
S1	0	−0.0352	0.0227	0.1208	−0.0871	0.0252	0.0006	−0.0312	0.0264	0.2379	−0.1658	0.0338	<.0001
S2	1	−0.5925	0.0768	<.0001	0.1236	0.0747	0.0978	−0.2004	0.0845	0.0178	−0.2607	0.1053	0.0133
Sex	0	0.5670	0.0245	<.0001	0.2115	0.0270	<.0001	−0.0503	0.0291	0.0840	−0.1375	0.0371	0.0002

추정된 odds ratio와 그 95% 신뢰구간은 [표 4]에 제시하였다.

[표 4] odds ratio추정치 및 신뢰구간

55 vs 57				
		odds ratio	95% 하한	95% 상한
Wage		1.005	1.002	1.007
J1	0 vs 2	0.955	0.876	1.042
J2	1 vs 2	1.400	1.285	1.526
S1	0 vs 2	0.965	0.923	1.009
S2	1 vs 2	0.553	0.476	0.643
Sex	0 vs 1	1.763	1.680	1.850

56 vs 57				
		odds ratio	95% 하한	95% 상한
Wage		1.000	0.998	1.003
J1	0 vs 2	1.097	0.998	1.207
J2	1 vs 2	1.053	0.959	1.158
S1	0 vs 2	0.917	0.872	0.963
S2	1 vs 2	1.132	0.977	1.310
Sex	0 vs 1	1.236	1.172	1.303

58 vs 57				
		odds ratio	95% 하한	95% 상한
Wage		1.001	0.998	1.004
J1	0 vs 2	0.978	0.885	1.081
J2	1 vs 2	1.025	0.928	1.133
S1	0 vs 2	0.969	0.920	1.021
S2	1 vs 2	0.818	0.693	0.966
Sex	0 vs 1	0.951	0.898	1.007

59 vs 57				
		odds ratio	95% 하한	95% 상한
Wage		1.005	1.001	1.008
J1	0 vs 2	0.829	0.732	0.940
J2	1 vs 2	1.018	0.900	1.151
S1	0 vs 2	0.847	0.793	0.905
S2	1 vs 2	0.771	0.627	0.947
Sex	0 vs 1	0.872	0.810	0.937

통계적으로 유의한 결과를 음영부분으로 표시하였으며 분석 결과를 살펴보면 55범주의 경우, 57범주에 대비하여 소득수준이 높으며, 가입종별이 임의 가입에 비해 지역가입인 경우가 많았고, 성별이 남성에 비해 여성인 성향을 보인 반면 소득활동여부는 대기에 비해 납부예외일 가능성이 더 낮았다.

56범주의 경우는, 57범주에 대비하여 성별이 남성에 비해 여성인 성향을 보인 반면에 소득활동여부는 대기에 비해 취득일 가능성이 더 낮았다.

58범주의 경우, 57범주에 대비하여 소득활동여부가 대기에 비해 납부예외일 가능성이 낮았다.

그리고 59범주의 경우는, 57범주에 대비하여 소득수준은 높은 반면에 가입종별은

임의가입에 비해 사업장 가입일 경우, 소득활동은 대기에 비해 취득이거나 납부예외일 경우, 그리고 성별은 남성에 비해 여성일 가능성이 낮았다.

이와 같이 다섯 범주에 대한 특성파악으로 기준범주 대비 개별 범주에 대한 설명변수의 유의성을 보이는 것은 가능하였으나, 결과에서 보는 바와 같이 기준 범주 57세보다 이른 수급 혹은 늦은 수급으로 갈수록 '일정한' 특성을 보이는 특징은 범주별로 양상이 일정하지 않음을 확인하였다. 따라서 수급 연령의 변화에 따른 특성파악을 위하여 3.2절에서는 기준 범주 57을 기준으로 57보다 이른 수급의 범주 그룹과 57세 보다 늦은 수급의 범주 그룹으로 나누어 세 그룹의 특징을 파악하여 보았다.

3.2 세 개의 군집을 이용한 다항 로짓 분석

57세 수급을 기준범주로 57세보다 이른 수급 범주인 55세와 56세 범주를 former group으로, 57세 이후 수급 범주인 58세와 59세 범주를 latter group으로 정의하였다. 종속 변수로 former, 57세, latter의 세 범주를 갖는 변수를, 설명변수로는 앞 절에서 사용한 변수들을 적용하였다. 세그룹에 대한 비례오즈모형가설 역시 기각되었기에 다항로짓분석을 시도하였으며 모형은 앞 절의 모형과 동일한 다항 로짓 모형으로 다음과 같다.

$$\log(\frac{\hat{\pi}_i}{\hat{\pi}_{57}}) = \beta_{i0} + \beta_{i1} \times Wage + \beta_{i2} \times J1 + \beta_{i3} \times J2 + \beta_{i4} \times S1 + \beta_{i5} \times S2 + \beta_{i6} \times Sex$$
$$i = \text{for}mer, latter$$

이를 통하여 추정된 모수는 [표 5]에 제시하였다.

[표 5] 군집을 이용한 다항 로짓 모형에 대한 β의 추정치

		기준범주(57)					
		former			latter		
		estimate	standard error	p-value	estimate	standard error	p-value
Intercept		0.7249	0.0562	<.0001	0.257	0.0637	<.0001
Wage		0.00297	0.00114	0.009	0.00236	0.00129	0.0676
J1	0	0.0144	0.0409	0.7242	-0.076	0.0464	0.1018
J2	1	0.2287	0.0405	<.0001	0.0231	0.046	0.6157
S1	0	-0.0543	0.0211	0.0099	-0.0747	0.0241	0.002
S2	1	-0.274	0.0663	<.0001	-0.2211	0.076	0.0036
Sex	0	0.4303	0.0228	<.0001	-0.0782	0.0265	0.0032

추정된 odds ratio와 그 95% 신뢰구간은 [표 6]와 같다.

[표 6] 군집의 odds ratio추정치 및 신뢰구간

former vs 기준범주(57)				
		odds ratio	95% 하한	95% 상한
Wage		1.003	1.001	1.005
J1	0 vs 2	1.015	0.936	1.099
J2	1 vs 2	1.257	1.161	1.361
S1	0 vs 2	0.947	0.909	0.987
S2	1 vs 2	0.760	0.668	0.866
Sex	0 vs 1	1.538	1.471	1.608

latter vs 기준범주(57)				
		odds ratio	95% 하한	95% 상한
Wage		1.002	1.000	1.005
J1	0 vs 2	0.927	0.846	1.015
J2	1 vs 2	1.023	0.935	1.120
S1	0 vs 2	0.928	0.885	0.973
S2	1 vs 2	0.802	0.691	0.930
Sex	0 vs 1	0.925	0.878	0.974

Odds ratio를 살펴보면 기준범주에 비해 former 그룹은 소득수준이 높으며, 가입종별이 임의가입에 비해 지역가입이 많다. 또한 대기에 비해 소득활동여부가 취득이거나 납부예외일 가능성은 낮았다. 그리고 성별이 남성에 비해 여성일 가능성은 높았다.

기준범주에 비해 latter 그룹역시 former 그룹과 마찬가지로 소득수준이 높고, 소득활동여부가 대기에 비해 취득이거나 납부예외일 가능성이 낮았다. 그리고 성별의 경우 여성에 비해 남성일 가능성이 높았다.

이로부터 기준 범주는 타 범주들과 구별되는 특성을 보임을 알 수 있었는데, 이러한 기준 범주의 특성으로 인하여 57세를 기준으로 former 그룹과 latter 그룹 간에는 소득수준, 가입종별, 소득활동여부, 성별 모두 유의한 요인이기는 하나 성별을 제외한 다른 요인들은 기준 범주 대비 former 그룹과 latter 그룹의 odds ratio값이 같은 방향(모두 1보다 크거나 모두 1보다 작은 값)을 보임으로 수급연령의 증감에 대하여 일정한 특성을 보이지 않음을 확인하였다. 이는 타 범주와 구별되는 기준범주의 특성으로 인한 결과로 3.3절에서는 former 그룹과 latter 그룹간의 특징을 살펴보기 위해서, 기준범주를 제거한 후 두 그룹에 대한 로지스틱 회귀 분석을 실시하였다.

3.3 former 그룹과 latter 그룹에 대한 로지스틱 회귀분석

본 절에서는 former 그룹과 latter 그룹 대한 특징을 살펴보기 위하여 기준범주를 제거한 후 former 그룹과 latter 그룹 대한 로지스틱 회귀 모형을 다음과 같이 세웠다.

$$\log(\frac{\widehat{\pi_{former}}}{\widehat{\pi_{latter}}}) = \beta_0 + \beta_1 \times Wage + \beta_2 \times J1 + \beta_3 \times J2 + \beta_4 \times S1 + \beta_5 \times S2 + \beta_6 \times Sex$$

두 그룹을 구별하는 요인들을 찾아내기 위하여 분석에서는 stepwise 변수선택법을 사용하여 적용하였고 그 결과 추정된 추정치와 odds ratio의 추정치와 95% 신뢰구간은 [표 7]에 제시되었다.

[표 7]로지스틱 회귀분석의 결과 β의 추정치와 odds ratio추정치 및 신뢰구간

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr
Intercept		1	0.4926	0.0361	185.6978	<.0001
J1	0	1	0.0859	0.0371	5.3558	0.0207
J2	1	1	0.2050	0.0368	31.0051	<.0001
Sex	0	1	0.5025	0.0165	927.2662	<.0001

former vs latter				
		odds ratio	95% 하한	95% 상한
J1	0 vs 2	1.090	1.013	1.172
J2	1 vs 2	1.227	1.142	1.319
Sex	0 vs 1	1.653	1.600	1.707

Odds ratio를 살펴보면 latter 그룹대비 former 그룹의 경우, 가입종별은 임의가입에 비해 사업장가입이거나 지역가입일 경우 그리고 성별은 남성에 비해 여성일 가능성이 높았다.

former 그룹과 latter 그룹에 대한 로지스틱 회귀분석을 통하여 성별과 가입종별이 그룹 간 일정한 특성을 보이는 유의한 요인임을 확인하였는데 성별의 유의성은 3.1과 3.2절의 다항 로짓 모형분석을 통해서도 파악 할 수 있었지만 가입종별의 일정한 특성은 기준 범주를 포함한 다항 로짓 분석으로는 파악할 수 없었던 요인이었다.

즉, 기준범주를 제거함으로써, former 그룹과 latter 그룹에 대한 로지스틱회귀분석을 통하여 57세보다 일찍 수급하는 그룹과 늦게 수급하는 그룹에 있어서 성별뿐 아니라 가입종별에 따른 차이가 통계적으로 유의함을 확인하였는데, 임의 가입에 비해 사

업장가입이거나 지역가입일 경우 이른 수급일 가능성이 높았으며 이로부터 가입종별은 이른 수급 혹은 늦은 수급에 대하여 일정한 추세를 보임을 추가적으로 확인하는 것이 가능하였다.

3.4 로지스틱 회귀분석에 의한 다섯 범주의 특성

3.1절과 3.2절의 다항로짓 분석 결과로부터 수급연령의 증감에 대하여 '일정한' 특성을 보이는 요인으로 성별만이 통계적으로 유의한 요인임을 확인하였다. 이는 다섯 범주 전체를 고려한 다항로짓 분석의 결과로, 기준이 되는 범주의 특성이 타 범주들과 달랐기 때문에 파악이 되지 않았던 '통계적으로 유의한 요인'을 추가적으로 확인하기 위하여 3.3 절에서는 기준범주를 제거한 후 former 그룹과 latter 그룹에 대한 특성분석을 시도하였다. 이에 따라 가입종별이 수급연령의 증감에 따라 '일정한' 특성을 보이는, 통계적으로 유의한 요인임을 추가적으로 확인하였다. 본 절에서는, 이러한 절차를 따라 파악한 가입종별의 특성에 대한 확인 과정으로 다섯 범주의 수급연령 각각에 대한 이산형 로지스틱 회귀분석을 실시하였다. 즉, 다섯 범주 각각에 대하여 다섯 개의 로지스틱회귀 모형을 설정하였는데 예컨대 55세 범주에 대한 로지스틱 회귀모형의 경우 55세에 조기노령연금을 수급하는 경우를 1로 그렇지 않은 경우를 0으로 보았다. 다섯 개의 로지스틱 회귀모형에 대한 분석 결과에 있어서 유의한 요인들과 이에 따라 도출된 각 범주의 특성은 [표 8]에 제시한 바와 같다.

[표 8] 다섯 개의 로지스틱 회귀 모형에 대한 변수선택 결과와 수급연령의 특성

범주	유의한 변수	특징
55세	소득수준, 가입종별, 성별, 소득활동 여부	높은 소득수준, 가입종별이 지역가입, 성별은 여성, 소득활동 여부는 취득이거나 대기
56세	소득수준, 가입종별, 소득활동 여부	낮은 소득수준, 가입종별이 사업장가입, 소득활동 여부는 납부예외이거나 대기
57세	소득수준, 가입종별, 소득활동여부, 성별	낮은 소득수준, 소득활동 여부가 취득이거나 납부예외, 성별은 남성, 가입종별은 지역가입
58세	가입종별, 성별	가입종별이 임의가입, 성별은 남성
59세	가입종별, 성별, 소득활동 여부	소득활동여부가 납부예외이거나 대기, 가입종별은 임의가입, 성별은 남성

3.4절에서 former 그룹과 latter 그룹에 대하여 유의한 요인으로 추가적으로 확인한 가입종별은 [표 8]에서 보는 바와 같이 다섯 범주에 대해 공통적으로 유의한 요인임을 알 수 있었다. 즉, 가입종별은 각각의 수급연령에 대하여 통계적으로 유의한 요인임을 확인하였다. 그러나 수급연령의 증감에 대한 가입종별의 통계적 유의성은 다항로짓 모형을 통하여 검정되어야 하는데 기준범주를 포함한 다항로짓 분석으로는 유의성을 확인할 수 없었던 사항이지만 기준 범주를 제거함으로써 former 그룹과 latter 그룹에 대한 로지스틱 회귀모형을 통하여 이러한 통계적 유의성을 밝힐 수 있었는데, 임의 가입에 비해 사업장가입이거나 지역가입일 경우 이른 수급일 가능성이 높음을

보였고 이는 위의 다섯 개의 로지스틱 회귀분석 결과가 55세에는 지역가입인 수급자가, 56세에는 사업장가입인 수급자가 많음을 보인 반면 58세와 59세의 경우에는 임의가입인 수급자가 많음을 보인 것과 같은 결과를 보임을 확인하였다.

비례오즈가정을 만족하지 않는 순위형 자료에 대하여 순위에 따른 설명요인의 일정한 특성을 파악하기위하여 대표 범주를 기준으로 하여 두 그룹으로 분류하고 이미 특성 파악이 이루어진 범주는 제거한 후 남은 범주들에 대하여 다항로짓 또는 로지스틱 회귀분석을 실시함으로써 순위에 따른 특성파악이 가능함을 보일 수 있었다.

4. 요약 및 결론

범주형 데이터 분석에 적용하는 로짓 모형의 경우 종속변수의 특성에 따라서 그 분석방법과 해석이 다르다. 본 연구에서는 비례오즈가정을 만족하지 않는 순위형 다범주 자료에 대하여 범주값의 증감에 따른 설명변수의 일정한 특성을 파악하기 위한 분석 방법으로 로짓모형을 적용하는 방법을 제시하였는데 범주가 많아 다항로짓 분석만으로 범주 값의 증감에 따른 설명요인의 일정한 추이 파악이 어려울 경우 필요에 따라 범주를 그룹화하거나 제거하는 방법을 적용할 수 있음을 보였다. 이에 대한 실증분석으로 국민연금 관리공단에서 주기적으로 모니터링하고 있는 가입종별, 성별, 소득수준, 소득활동여부를 설명요인으로 한 조기노령연금의 수급개시연령의 특성파악을 위한 분석 결과와 이에 대한 해석을 예시하였다.

종속변수로 사용된 조기노령연금의 수급개시연령은 55부터 59까지의 다섯 개의 범주를 갖고 있는데, 셋 이상의 범주를 갖는 순위형 종속변수의 경우 비례오즈 모형을 적용할 수 있으나 본 연구에서 다룬 데이터의 경우 비례오즈 가정을 만족하지 않았기에 다항 로짓 모형을 이용하여 분석하였다.

이러한 다항 로짓 모형의 경우 기준범주를 정하여 각 범주에 속하는 특징들을 파악할 수 있다. 본 연구에서는 55에서 59까지의 대표 값인 57범주를 기준으로 하여 55, 56, 58, 59에 속하는 특징들을 파악하였고 그 결과 57범주 대비 각 범주에 속하는 특징들을 파악할 수는 있었지만, 수급연령의 증감에 대한 요인들의 일정한 특성을 파악하는 데는 어려움이 있었다.

이에 대한 해결책으로 57세를 기준범주로 정하여 수급연령의 많고 적음에 따른 특징을 파악하기를 시도하였고 수급개시연령 55, 56을 former 그룹으로 58, 59를 latter 그룹으로 묶어서 세 범주에 대한 다항 로짓 분석을 실시하였다. 하지만 기준범주를 대비한 former 그룹과 latter 그룹에 대한 특성을 파악하기가 어려웠다. 이는 기준범주인 57 범주가 former 그룹과 latter 그룹에 비해서 특성이 뚜렷하게 다르기 때문임을 파악하였다.

따라서 이른 수급 혹은 늦은 수급을 선택함에 따른 요인들의 일정한 성향을 파악하기 위하여 앞에서 이미 특성을 파악한 기준범주를 제외한 former 그룹과 latter 그룹에 대한 로지스틱회귀분석을 실시하였다. 이 과정을 통하여 앞선 분석에서는 파악할 수 없었던, 이른 수급과 늦은 수급에 대해 일정한 특성을 보이는 유의한 요인을 찾을 수 있었다. 두 그룹은 가입종별과 성별에 있어 유의한 차이를 보였는데 former 그룹의 경우 가입종별이 사업장가입이거나 지역가입일 가능성이 높았으며, 여성일 경우 이에 속하는 가능성이 높았다. 반면 latter 그룹은 임의가입이거나 남성일 경우가 이에

속해 두 그룹간의 특성을 구별할 수 있었다. '가입종별'의 특성으로 기준범주를 제거함으로써 얻을 수 있었던 통계적 유의성과 수급연령의 증감에 대하여 보이는 일정한 특성에 대한 확인과정으로 다섯 범주의 수급연령 각각에 대하여 설정한 로지스틱 회귀모형으로부터 가입종별은 모든 범주에 대하여 공통적으로 유의한 요인임을 확인하였다. 이로부터 가입종별은 각각의 수급연령에 대하여 유의한 요인이며 이른 수급과 늦은 수급의 선택에 있어서도 통계적으로 유의한 요인일 뿐 아니라 선택하는 수급연령의 증감에 있어서 일정한 특성을 보임을 알 수 있었다.

이와 같은 분석 결과는, 조기노령연금의 관리에 있어서 이른 수급과 늦은 수급의 대상자를 예측 혹은 분류가 가능하도록 도움으로써 수급자 관리에 필요한 정보를 제공할 수 있다. 또한 설명요인의 값의 변화에 따른 수급연령의 증감에 대한 가능성의 비율로써 odds ratio 값은 조기노령연금의 재정 관리에 있어서 이른 수급과 늦은 수급간의 현실적인 차이를 반영하기 위한 기준값으로 사용할 수 있다.

본 연구에서는 비례오즈가정을 만족하지 않는 순위형 다범주 자료의 분석에 있어서 범주값의 증감에 따른 설명변수의 일정한 특성을 파악하기 위한 분석 방법을 제안하였다. 순위형 다범주값의 증감에 따른 설명 요인의 일정한 특성을 파악하는데 있어서 다항로짓 모형만으로는 파악하기 어려운 경우, 기준범주 대비 범주값이 증가한 그룹과 감소한 그룹으로 범주를 그룹화한 후 그룹화 된 범주의 특성을 파악하고, 기준 범주와 나머지 범주간의 특성을 살펴봄으로써 기준 범주의 특성을 파악하고 (3.2절), 기준범주를 제외한 두 그룹의 특성을 로지스틱회귀모형을 통하여 파악함으로써 순위형 범주값의 증감에 대해 통계적으로 유의한 설명요인들을 추가적으로 찾을 수 있음을 설명하였다 (3.3절). 3.3절의 결과에 대한 확인과정으로 수급연령의 다섯 범주 각각에 대한 로지스틱회귀 모형을 적합하여 분석함으로써 가입종별은 각 수급연령에 대하여 유의한 요인일 뿐만 아니라 이른 수급과 늦은 수급의 선택에 있어서도 일정한 특성을 보이는 요인이며 통계적으로 유의한 요인임을 보였다 (3.4절).

본 연구는, 비례오즈 가정사항을 만족하지 않는 순위형 다범주 자료를 분석하는 방법에 있어서 조기노령연금의 수급개시 연령에 따른 특성 분석을 예시로 보였으며, 자료가 제공한 네 개의 설명변수만을 근거로 유의한 요인을 찾기를 시도하였으나 더 많은 요인들이 주어진다면 수급개시연령에 따른 조기노령연금의 특성파악이 좀 더 구체적으로 이루어 질 뿐만 아리라 이른 수급과 늦은 수급에 대한 분류에 있어서 정확도가 높은 모형을 제시할 수 있으리라 사료된다.

참고문헌

1. 강현철, 한상태, 최종후, 이성건, 김은석, 엄익현, 김미경 (2006). 고객관계관리를 위한 데이터마이닝 방법론, 자유아카데미, 경기.
2. 국민연금관리공단 (2007). 2006년 국민연금 통계연보, 19, http://www.nps.or.kr/
3. 박용규 (2001). 통계시리즈(XVIII): 범주형자료의 분석, 가정의학회지, 22(5), 631-644
4. 이선우 (2001). 장애인의 경제활동유형 결정요인에 대한 연구: Multinomial Logit 을 이용한 분석, 사회 복지연구, 18, 113-134

5. 이영섭, 박주완 (2007). 기업 인적자원 관련 변수를 이용한 기업 신용점수 모형 구축에 관한 연구, 응용통계연구, 20(3), 423-440
6. Agresti, A.,(1996) *An Introduction to Categorical Data Analysis* . Wiley Series in Probability and Statistics.
7. Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Economics*. Cambridge university Press.
8. Neter, J., Wasserman W., & Kutner, M. (1990) *Applied Linear Statistical Models*, Richard D. IRWIN, Inc.
9. SAS (2000). *SAS System for Windows* V.9.1, SAS Institute Inc.

[접수일(2008년 5월 22일), 수정일(2008년 6월 19일), 게재확정일(2008년 7월 7일)]