응용통계연구 제 **21**권 2호, 2008년, pp. 183–200

로지스틱 회귀모형을 이용한 유족연금 수급 분석*

김미정[1] 김진형[2]

요 약

국민연금의 효율적인 운영을 위하여 고령화, 저출산과 같은 사회현상에 대비한 연금관리를 위한 연구가 요구되고 있다. 본 연구는 유족연금의 발생을 예측하고 유족연금의 발생가능성 정도에 따라 대상자들을 분류하기 위한 통계적 모델을 제안하기 위하여 두 단계의 로지스틱 분석을 실시하였다. 첫 단계의 분석으로부터, 전체 대상자에 대하여 유족연금의 발생에 영향을 주는 주요인의 특성과 국민연금의 종류를 파악하고 이를 대상으로 유족연금의 발생에 대한 로지스틱 회귀모형을 적용하되 대상자를 합리적으로 등급화하기위한 모델을 제안하고 이를 일반적인 로지스틱모델과 비교하였다. 정확도, 민감도, 특이도와 사후 확률의 분포를 비교하고 K-S통계량을 통하여 등급의 타당성 평가와 리프트 그래프를 통한 모델의 예측력평가를 함으로써 합리적 등급분류를 통한 대상자관리가 가능한 통계적 모델임을 보였다. 예측된 통계적 모델을 적용하여 유족연금 수급유무와 등급별 분류, 등급에 따른 유족연금액 예측을 통하여 효율적인 연금관리 방안을 제안할 수 있다.

주요용어: 국민연금, 유족연금, 로지스틱 회귀분석, 정확도, 타당성 평가.

1. 서론

국민연금은 노령으로 인하여 소득상실의 위험에 직면한 노인들에게 일정한 소득을 제공하여 안정된 노후생활을 영위할 수 있게 하는 공적 연금제도로 1988년 상시 근로자 10인 이상 사업자 근로자를 대상으로 시작하였다. 이렇게 시작된 국민연금은 크게 노령연금, 장애연금, 유족연금으로 나눌 수 있다.

노령연금은, 가입기간, 연령, 소득활동 유무에 따라서 수급요건과 급여수준이 달라지는 연금으로 완전, 감액, 재직자, 조기, 특례노령연금으로 구분되며, 국민연금이 생긴 연유에 가장 근접한 연금이다. 장애연금은 질병 또는 부상의 원인으로 노동능력이 상실 또는 감소된 경우에 생계안정을 위하여 지급되는 급여로 장애 정도에 따라 지급이 된다. 유족연금은 가입자 또는 10년 이상 가입했던 이력을 가진 자나 노령연금 또는 장애등급 2급 이상의 장애연금 수급권자가 사망한 경우에 그 유족의 생활을 보장하기 위하여 배우자, 자녀, 손자, 조부모에게 지급되는 연금을 말한다.

* 본 연구의 자료는 국민연금관리공단의 공동연구 자료임.

1) (120-749) 교신저자. 서울시 서대문구 신촌동 134, 연세대학교, 수리 과학 연구소, 연구교수.
E-mail: mjkim@yonsei.ac.kr
2) (120-749) 서울시 서대문구 신촌동 134, 연세대학교, 산업공학과, 석사과정.
E-mail: statkjh@yonsei.ac.kr

유족연금은 다른 연금과는 달리 그 발생이 연금의 종류에 따라 다양하게 나타나며, 유족연금의 수급요건을 충족시키지 못했을 경우 사망일시금을 지급하거나 연금 지급이 종료된다. 노령연금의 경우에는 발생여부가 가입자의 나이에 따라 결정되며 장애연금의 경우에는 가입자가 질병 또는 부상을 입어 노동 능력을 상실한 경우 발생한다. 반면 유족연금은 국민연금 가입자 또는 10년 이상 가입했던 이력을 가진 자나 노령연금 수급자 또는 장애연금 수급권자가 사망하는 경우에 그 유족에게 지급되기 시작하며 이러한 연금의 발생을 예측하는 것은 연금을 효율적으로 관리하는데 있어 중요하다.

본 연구에서는 연금의 발생이 가입자 및 수급권자의 상태에 따라 다양하게 나타나고 있는 유족연금을 중심으로 연금의 발생에 영향을 미치는 요인들을 파악하고 유족 연금 발생 여부를 예측하는 모델을 로지스틱 회귀모형을 적용하여 제안한다.

이를 위하여 첫째로는 유족연금 및 일시금 지급의 전체 대상자들 가운데 사망자를 대상으로 유족연금의 발생에 영향을 미치는 요인과 그 특성을 파악하고, 이들 가운데 유족연금의 발생에 가장 많은 영향을 끼치는 그룹을 중심으로 유족연금 발생에 대한 통계적 예측 모델을 제시하되 유족연금발생 확률의 크기에 따라 등급분류가 가능하도록 적합한 로지스틱 회귀 모형을 제안한다. 유족연금 및 일시금 지급의 전체 대상자란, 국민 연금 '가입자 및 대기자'와 '장애연금' 수급자 그리고 '조기노령연금' 수급자를 말하며, 국민 연금 '가입자 및 대기자'에서 가입자란 국민 연금에 가입한 자로써 아직 장애연금이나 노령연금의 수혜를 받고 있지 않는 자를 의미하며, 대기자란 국민연금 가입이력을 가지고 있는 잠재적 자격자를 의미한다. 본문에서는 편의상 '가입자 및 대기자', '조기노령연금', '장애연금'을 국민연금의 종류로 명명하기로 한다.

첫 번째 제시되는 모형을 통해서는 유족연금 및 일시금 수급 대상자 전체의 특성을 파악하고, 두 번째 제시되는 모형에서는 유족연금 발생에 가장 많은 영향을 미치는 연금의 종류를 중심으로 유족 연금 발생에 대한 예측모형을 세움으로써 유족 연금관리 방안의 틀을 한 가지로 제시하고자 한다. 이렇게 함으로써, 연금의 종류에 따른 연금관리 정책을 각각 제시하고 관리해야하는 불편함과 그에 따른 비용을 줄일 수 있다고 생각한다. 그러나 필요에 따라서는 특정 연금의 종류에 따른 정책이 요구될 수도 있으므로 이에 대한 모형을 4장에서 다루었다.

둘째, 효율적인 유족연금 관리 방안을 위해 제안한 모델을 일반적인 두 개의 로지스틱 모델과 비교하였다. 설명 변수들의 주효과만을 고려한 모델 (1), 주효과와 교호작용을 고려한 모델 (2) 그리고 본 연구에서 제안하는 모델로 최적의 자료를 대상으로 로지스틱 회귀식을 적용한 모델 (3)의 세 모형 중 모델 (3)이 효율적인 연금관리를 위해 합리적이고 적합도가 높은 예측 모델임을 보였다. 이를 위해 대상자들의 유족연금 수급 가능성의 정도에 따른 분류를 위한 기준점으로 제시 될 threshold 값에 따른 정확도, 민감도(sensitivity), 특이도(specificity)와 사후 확률의 분포를 비교하였고 모형의 평가를 위하여 세 모형의 등급에 따른 반응률을 비교하고 모델에 따른 평점표의 타당성 평가를 위하여 Kolmogorov-Smirnov 통계량을 비교하였다.

이렇게 제시된 모델로 각 대상자의 유족연금 발생에 대한 확률적 예측이 가능하며 유족연금의 발생에 영향을 주는 요인인 사망연령, 가입기간, 가입종별, 소득수준 그리고 성별

에 따른 등급별 분류를 통하여 유족연금 발생에 대비한 연금관리 방안을 등급별 특성에 맞게 제안할 수 있다. 향후 발생할 유족연금 대상자를 예측하고 이들을 유족연금 발생 확률에 따라 등급화하고 등급에 따라 분류된 그룹의 특성을 파악함으로써 그룹별 유족 연금액의 합리적인 예측이 가능하고 따라서 고령화와 저출산과 같은 사회적 특성을 반영하지 못하여 발생할 수 있는 비적절한 예측으로 인한 손실을 줄임으로 보다 현실적인 연금운영을 통해 재정적자와 같은 문제를 막을 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 국민연금과 관련한 문헌을 살펴본 후, 분석자료를 대상으로 유족연금에 영향을 주는 요인들을 파악하였다. 3장에서는 '가입자 및 대기자'그룹의 사망자가운데 유족연금의 발생에 대한 통계적 모델로 본 연구에서 제안하는 모델과 두 개의 일반적인 로지스틱 모델을 비교하였다. 4장에서는 '조기노령연금'과 '장애연금' 수급자를 대상으로 유족연금의 발생에 대한 통계적 모델을 본 연구에서 제안하는 모형으로 적합하였으며, 5장에서는 분석한 자료를 근거로 유족연금 대상자의 등급별 분류 예시를, 마지막으로 6장에는 결론 및 향후 연구방향에 대해 기술하였다.

2. 유족연금의 요인분석

김지훈 (2000)은 공무원연금에 대한 재정전망에 대해 알아보기 위하여 보험수리적 분석방법을 적용하여 공무원 연금의 재정을 전망하였고 안홍순 (2007)은 국민연금 재정안정화를 위해서 표준소득월액과 전체가입자의 평균소득월액이 가입자의 소득 증대에 탄력적으로 적용될 수 있는 연금자동안정장치를 도입하여 노인부양비율의 상승으로 인한 재정 부담을 완화하는 기본연금산식의 구조개혁 그리고 재정방식을 부과방식으로 전환할 것을 제안하고 있다. 이근홍 (2006)은 국민연금의 저부담·고급여의 구조를 유지함에 따라 연금재정이 불안정할 것으로 예측하여 재정의 안정화를 위해 연금의 기여율을 인상하고 보험료의 징수율을 조정할 필요가 있다고 했다. Bergh 등 (2007)은 다중 로짓분석을 이용하여 5년 이상 의료보험에 가입한 사람들에 대하여 병이 자주 걸리는 가입자와 일반 가입자 중 장기환자와 장애연금을 받는 사람들에 대한 예측요인을 분석하였다. Humphreys 등 (2007)은 미국의 남북전쟁의 퇴역병사들에 대한 연금데이터를 바탕으로 하여 인종에 따른 당뇨병의 차이를 비교하기 위하여 로짓모형을 적용하여 분석했으며 인종에 따른 당뇨병의 발병에는 차이가 없으며 소득수준과 육체적인 노동에 대한 차이만 있음을 보였다. Huberman 등 (2007)은 확정기여연금의 수여자와 공급자의 비율을 결정하기 위하여 프로빗 모델과 Tobit모델을 제안하였다.

본 연구에서는 국민연금 자료에 로지스틱 회귀식(logistic regression)을 적용하여 유족연금의 발생에 대한 통계적 모델을 세우고 연금재정의 리스크 관리를 위한 시스템을 제안한다. 자료의 분석은 모두 SAS (2000) V.9.1 을 사용하였다.

일반적으로 설명변수가 p개인 로지스틱 모형은 다음과 같다.

$$\log\left(\frac{\Pr(Y=1|X_1,\ldots,X_p)}{\Pr(Y=0|X_1,\ldots,X_p)}\right)=\beta_0+\beta_1x_1+\cdots+\beta_px_p,$$

표 2.1: 모형설정에 사용된 변수

변수	정의	특성
유족연금	유족연금의 여부를 나타내는 종속변수	1=유족연금, 0=일시금 또는 지급종료
성별	성별을 나타내는 가변수	1=남성, 0=여성
기간	국민연금의 가입기간	가입개월 수
소득수준	소득수준	1부터 45까지의 연속형 자료
종류	연금의 종류	1=가입자 및 대기자 2=장애연금 3=조기 노령연금
종별	가입종별	0=사업장가입 1=임의가입 2=지역가입
나이	사망나이	사망시점의 연령

여기서, $Y = 1$은 유족연금의 발생을 나타내며 로지스틱 모형의 결과로부터 얻는 확률은 p개의 설명 변수값이 주어졌을 때 유족연금이 발생할 확률로 다음과 같이 표현된다.

$$\Pr(Y = 1|X_1, \ldots, X_p) = \frac{\exp\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)}{1 + \exp\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)}. \tag{2.1}$$

유족연금 발생에 영향을 주는 요인들의 특성을 분석하기 위하여 1988년 1월부터 2007년 5월까지의 국민연금 '가입자 및 대기자', '장애연금' 수급자 그리고 '조기노령연금' 수급자 가운데 사망자 451,011명을 분석 대상으로 하였다. 이들은 모두 유족연금 또는 반환 일시금을 받았거나 사망 일시금 지급 또는 연금지급이 종료된 사람들이다. 여기서 반환 일시금이란, 대기자, 즉 국민연금 가입이력을 가진 잠재적 자격자가 10년 이상의 가입 기간을 채우지 않았을 경우 가입자가 사망했을 때 유족에게 일시급으로 지급하는 것으로 유족이 없는 경우 지급되는 사망 일시금과는 구별되므로 유족 연금의 발생으로 보았다. 유족연금의 발생여부는, '가입자 및 대기자'의 경우 수급권자가 유족연금 또는 일시 반환금을 받은 경우 1의 값을, 사망 일시금을 받은 경우 0의 값을 갖는 것으로 되어 있으며, '장애 연금'과 '조기 노령연금'의 경우 유족 연금을 받은 경우 1 그렇지 않은 경우 0의 값을 갖는 것으로 되어 있다. 이를 유족 연금 발생여부를 나타내는 종속 변수로 사용하였으며, 사용된 설명변수들은 국민 연금 수급자관리에 있어서 고려하는 모든 설명 변수를 포함한 것으로 성별, 가입기간, 소득수준, 사망나이, 연금의 종류 그리고 가입종별이며 이에 대한 설명은 표 2.1에 제시하였다.

데이터를 살펴보면, 성별의 경우 남성이 88.74%이며, 여성은 11.23%이다. 가입종별은 연금이 지급되기 직전의 가입종별로 사업장가입이 36.57%이고, 임의가입은 11.22%, 지역가입은 52.21%이며, 연금의 종류는 수급자들의 연금가입상태로, 가입자 및 대기자는 76.17%, 장애연금은 4.06% 그리고 조기노령연금은 19.77%였다. 여기서 가입기간은 가입자가 연금가입 후 연금이 수급될 때까지의 기간을 개월 수로 나타냈으며, 소득수준은 소득이 있는 경우의 표준보수액을 기준으로 하여 1부터 45등급의 소득수준을 나타낸다. 1988년 1월 이후부터 95년 4월 까지는 53등급으로 나누던 것을 95년 4월 이후에는 45등급으로 조정되어 분석에서는 95년 4월까지의 등급을 45로 보정하여 평균 소득수준을 계산하였다. 보정은 53등

표 2.2: 자료 요약

변수	자료수	평균값	표준편차	최소값	최대값
기간	451011	59.35519	45.01422	0	232
소득수준	451011	19.33669	12.46583	0.956121	45
나이	451011	51.26480	11.56650	16	78

표 2.3: 추정치와 odds ratio의 95% 신뢰구간

Parameter		Estimate	odds ratio	신뢰 하한	신뢰 상한	Standard Error	Wald Chi-Square	Pr
Intercept		−4.2725	-	-	-	0.0365	13732.64	< .0001
나이		0.0445	1.046	1.044	1.047	0.0005	7874.854	< .0001
기간		0.0013	1.001	1.001	1.002	0.0001	111.8141	< .0001
소득수준		0.0331	1.034	1.033	1.035	0.0005	4862.688	< .0001
종별	0 *vs* 2	−0.0366	0.964	0.941	0.987	0.0122	8.9801	0.0027
종별	1 *vs* 2	0.0784	1.082	1.049	1.116	0.0158	24.5445	< .0001
성별	1 *vs* 0	2.3388	10.369	10.148	10.595	0.0110	45286.19	< .0001
종류	1 *vs* 3	1.5494	4.709	4.569	4.852	0.0153	10242.48	< .0001
종류	2 *vs* 3	−0.8184	0.441	0.424	0.459	0.0198	1714.727	< .0001

급으로 나누었을 때의 53등급과 45등으로 나누었을 때의 45등급이 동일한 등급임을 가정하였으며 1995년 4월까지의 소득수준에 대하여 linear interpolation을 하였다. 사망나이의 경우 사망에 의해 연금이 정지된 때의 나이를 의미하며, 이에 대한 요약은 표 2.2와 같다.

로지스틱 회귀분석의 결과를 통하여 나온 추정치와 odds ratio를 살펴보면, 유족연금의 발생에 영향을 미치는 요인들은 표 2.3의 결과와 같다.

전체 유족연금의 발생에 영향을 미치는 요인들로는 연속형 변수인 경우, 사망연령이 높을수록, 가입기간이 길수록 소득수준이 증가할수록 유족연금이 발생할 가능성이 높았다. 또한 범주형의 변수를 해석해 보면, 가입종별이 지역가입에 비해 사업장가입일 경우 유족연금의 발생 원인이 장애연금인 경우 조기노령연금에 비해 유족연금일 가능성이 낮았다. 반면에 임의가입인 경우 지역가입에 비해 유족연금의 발생가능성이 높았고, 남성이 여성에 비해 유족연금의 발생가능성이 높았으며, 가입자 및 대가지인 경우 조기노령연금에 비해 유족연금의 발생가능성이 높았다.

이를 통하여 유족연금의 발생에 가장 큰 영향을 미치는 요인은 성별로 여성에 비해 남성이 유족연금일 가능성은 10.369배 높았고 다음으로 '가입자 및 대기자'인 경우가 '조기노령연금'에 비해 4.709배 유족연금일 가능성이 높았다. 또한 유족연금에 가장 낮은 영향을 미치는 요인은 '장애연금'이 '조기노령연금'에 비해 0.441배 낮았으며, 다음으로 사업장가입이 지역가입에 비해 0.964배 낮았다.

본 연구의 관심사항으로 유족연금의 발생에 가장 많은 영향을 미치는 요인을 찾아보면, '가입자 및 대기자'가 '조기노령연금'에 비해 유족연금의 발생에 더 많은 영향을 미쳤으며, '장애연금'이 '조기노령연금'에 비해 유족연금의 발생에 덜 영향을 미쳤다. 따라서 '가입자

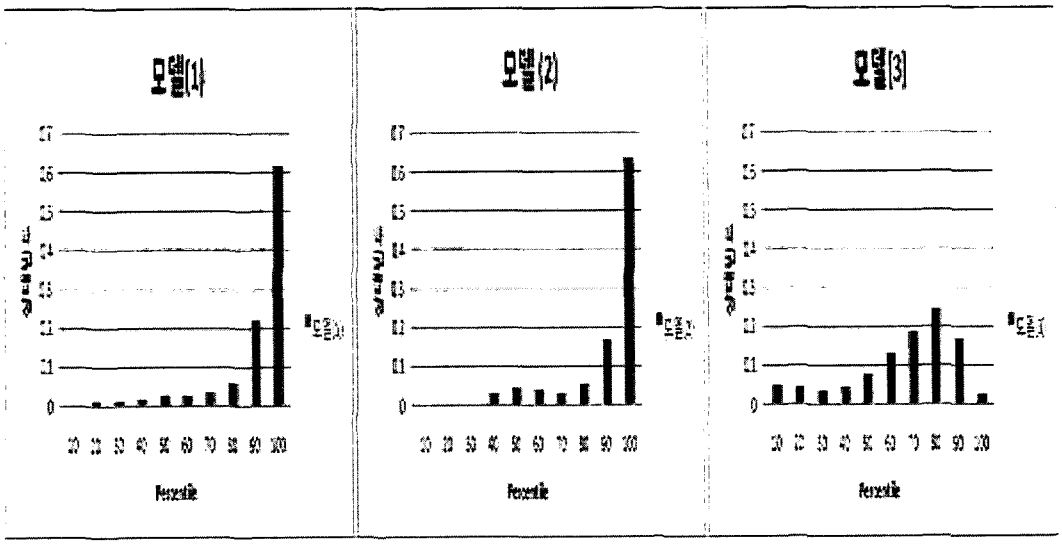


그림 3.1: 전체 자료에 대한 세 모델의 사후 확률 분포

및 대기자'를 유족연금의 발생에 가장 많은 영향을 미치고 있는 그룹으로 파악하였다. 표 2.3에 설명한 설명변수들은 모두 유의수준 0.05에서 유의하며, 결과는 다음 단계의 로지스틱모형 설정에 적용되었다.

3. 유족연금의 예측모델

유족연금의 예측모델을 제안하기 위하여 '가입자 및 대기자' 가운데 사망자 343,537명 중 성별이 불분명한 9개의 결측치를 제거한 343,528명의 자료를 분석대상으로 하였다. 설명변수는 사망나이, 가입기간, 소득수준, 가입종별 그리고 성별이며 다중 공선성의 문제를 제거하기 위해서 step-wise selection을 적용하였다. 자료의 특성을 살펴보면, 반응변수는 0 또는 1의 값을 가지며, 1의 값을 갖는 자료가 전체의 86.27%로 구성되어 있어서 통계적 모델 설정 후 예측된 사후 확 률의 분포가 오른편에 지나치게 편중되어 왼편 꼬리가 길어지는 경향을 보임으로 대상자의 유족연금 수급여부의 예측 확률에 따른 등급분류의 기준점을 결정하기가 어려웠다 (그림 3.1 참고). 등급분류가 제대로 되려면, 등급분류의 기준이 되는 사후 확률값이 등급에 따라 차별화되고 등급이 높아질수록 유족 연금지급의 비율은 높아지고 일시금 지급의 비율은 낮아지도록 순위화가 되어야 하기 때문이다.

이러한 문제점을 해결하기 위해서, 본 연구에서 제시하는 모형에서는, 주어진 자료에서 최적의 자료를 추출하여 모델을 설정하고 이를 일반적인 모델과 비교하였다. 본 자료의 경우, 분석 대상 자료의 두 반응범주의 빈도수를 동일하게 맞춘 것을 최적의 자료로 보았다. 즉, 제안하는 모형의 경우, 높은 빈도를 갖는 범주의 자료로부터 낮은 빈도를 갖는 범주의 빈도 수 만큼의 자료를 임의로 추출한 후 두 범주의 자료에 대하여 로지스틱 모델을 적용

표 3.1: 분석 대상 빈도

유족연금 발생여부	모형 (1)과 (2)에서 고려한 자료의 전체 빈도수	모형 (3)에 고려한 자료의 전체 빈도수
1 = 유족연금 또는 일시 반환금	각 296,366 명	47,162 명
0 = 사망 일시금	각 47,162 명	47,162 명

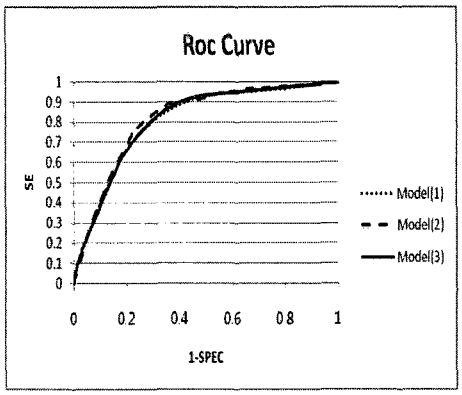


그림 3.2: 세모형의 ROC curve

하였다.

데이터의 주효과만을 고려한 모형 (1), 주효과와 교호작용효과를 고려한 모형 (2) 그리고 본 연구에서 제안하는 모형으로, 두 반응변수의 빈도수, 즉 일시금지급 수급자와 유족연금 수급자의 빈도수를 동일하게 맞추어 추정한 모형 (3)을 비교하며 분석대상 자료에 대한 요약은 표 3.1와 같다.

고려대상 자료의 70%를 training 데이터로 사용하여 세 모형을 적합하였으며, 30%의 자료는 validation 자료로 사용하였다. 표 3.2는 이에 대한 로지스틱 회귀 분석 결과를 보여주고 있다.

세 모형에 대한 예측 검정력을 비교하기 위하여, 반응변수의 관측값과 예측 확률값간 최대 상관관계값이 1이 되도록 보정한 Max rescaled R-square를 살펴보면, 모형 (1), (2), (3) 각각의 경우 27.16%, 31.23%, 36.87%로 모형 (3)의 값이 모형 (1)과 (2)보다 각각 35.18%와 18.0%만큼씩 증가했다. 세 모형에서 Max rescaled R-square 값이 비교적 높지 않은 이유는 위에서 언급한 다섯 개의 기본 변수에 따른 유족연금수급의 특성을 파악하고자 한 자료의 특성에 의한 결과이다.

세 모형의 ROC(Receiver Operation Characteristic) curve는 그림 3.2과 같다.

세 ROC curve가 유사한 모양을 보임으로 threshold에 상관없이 세 모형의 정확도에는 큰 차이가 없음을 확인할 수 있다. 즉, 유족연금 수급과 일시금 수급에 대한 분류를 옳게 의사 결정할 확률은 세 모형 모두 비슷하다고 볼 수 있다. 그러나 민감도와 특이도는 유족연금과 일시금수급 여부에 대한 분류의 기준점으로 제시될 threshold에 따라 결정되고 주어지는 threshold에 따른 민감도와 특이도는 각 모형의 ROC curve마다 그 값들이 서로 다르

표 3.2: 세 모형의 로지스틱 회귀분석의 결과

주효과만을 고려한 모델(1)

Parameter		Estimate	Standard Error	Wald Chi-Square	Pr
Intercept		−3.0219	0.0554	2978.261	< .0001
나이		0.0481	0.0007	5255.388	< .0001
기간		0.0025	0.0002	218.7984	< .0001
종별	0 *vs* 2	−0.1617	0.0417	15.0301	0.0001
	1 *vs* 2	−0.0589	0.0394	2.2354	0.1349
소득수준		0.0416	0.0007	3517.635	< .0001
성별	1 *vs* 0	2.3824	0.0156	23451.68	< .0001

주효과와 교호작용을 고려한 모델(2)

Parameter		Estimate	Standard Error	Wald Chi-Square	Pr
Intercept		0.6267	0.0539	135.0095	< .0001
성별	1 *vs* 0	−1.6617	0.0585	806.6531	< .0001
종별	0 *vs* 2	−2.1188	0.0613	1192.821	< .0001
종별	1 *vs* 2	2.8362	0.4198	45.6468	< .0001
기간		−0.0025	0.0004	52.0184	< .0001
소득수준		0.0158	0.0013	158.9891	< .0001
나이		−0.0215	0.0011	384.516	< .0001
성별*종별	1*0 *vs* 0*2	−0.4295	0.0329	170.1253	< .0001
성별*종별	1*1 *vs* 0*2	−0.6451	0.0704	83.9702	< .0001
기간*성별	1 *vs* 0	0.0064	0.0004	322.5323	< .0001
소득수준*성별	1 *vs* 0	0.0298	0.0014	458.3585	< .0001
나이*성별	1 *vs* 0	0.0766	0.0012	4111.37	< .0001
나이*종별	0 *vs* 2	0.0470	0.0012	1561.462	< .0001
나이*종별	1 *vs* 2	−0.0320	0.0066	23.2467	< .0001
기간*종별	0 *vs* 2	−0.0031	0.0003	95.138	< .0001
기간*종별	1 *vs* 2	0.0034	0.0016	4.4706	0.0345
소득수준*종별	0 *vs* 2	0.0243	0.0013	330.1218	< .0001
소득수준*종별	1 *vs* 2	−0.0166	0.0037	20.7718	< .0001

반응변수에 대한 빈도수를 동일하게 맞춘 모델(3)

Parameter		Estimate	Standard Error	Wald Chi-Square	Pr
Intercept		−5.2830	0.0798	4379.319	< .0001
나이		0.0517	0.0009	3244.086	< .0001
기간		0.0019	0.0002	69.193	< .0001
종별	0 *vs* 2	−0.2537	0.0584	18.8824	< .0001
종별	1 *vs* 2	−0.0704	0.0550	1.6362	0.2009
소득수준		0.0474	0.0010	2434.751	< .0001
성별	1 *vs* 0	2.6852	0.0282	9055.604	< .0001

표 3.3: 세 모형의 예측 검정력

Max rescaled R-square	
모델 (1)	0.2761
모델 (2)	0.3123
모델 (3)	0.3687

표 3.4: 세 모델의 분류표 (training 데이터의 결과)

Threshold	모형 (1)			모형 (2)			모형 (3)		
	정확도	민감도	특이도	정확도	민감도	특이도	정확도	민감도	특이도
0.2	86.6	99.7	4.4	86.3	100	0.0	66.0	95.2	36.9
0.3	86.6	99.1	8.3	86.3	99.9	0.7	71.9	93.3	50.5
0.4	86.3	98.0	12.9	87.1	98.8	13.9	74.7	89.4	59.9
0.5	86.3	96.7	21.2	87.7	96.6	31.6	75.6	82.1	69.2
0.6	86.7	95.5	31.7	87.7	94.3	46.6	73.6	68.4	78.9
0.7	87.4	94.0	45.7	87.5	92.3	57.4	67.5	48.0	87.0
0.8	85.5	89.7	58.8	84.4	87.5	65.1	58.6	21.6	95.6
0.9	69.3	67.8	78.9	71.1	70.1	77.2	51.3	2.9	99.7
1.0	13.8	0.0	100	13.7	0.0	100	50.0	0.0	100

	threshold	정확도	민감도	특이도
모형 (1)	0.86	79.9	81.6	69.3
모형 (2)	0.86	78.9	80.2	70.9
모형 (3)	0.50	75.6	82.1	69.2

표 3.5: 전체자료에 대한 분류표 (괄호안은 validation자료에 대한 값)

	threshold	정확도	민감도	특이도
모형 (1)	0.86	80.70 (80.04)	82.67 (81.76)	68.37 (69.20)
모형 (2)	0.86	78.96 (79.04)	80.25 (80.35)	70.85 (70.74)
모형 (3)	0.50	80.78 (75.64)	82.79 (82.20)	68.19 (69.09)

다. 즉, 세 ROC curve의 모양이 같더라도 원하는 민감도와 특이도를 얻기 위하여 요구되는 threshold의 값은 각각의 ROC curve마다 크게 다르게 나타난다. threshold에 따른 세 모형의 민감도, 특이도, 정확도를 비교하기 위하여 70%의 training 데이터에 적용한 분류표를 표 3.4에 보였다. 세 모형에 대한 평가로 전체자료와 validation자료에 대하여 정확도, 민감도, 특이도를 표 3.5에 비교하였으며, 전체 자료에 대하여 사후 확률의 분포를 비교하였다 (그림 3.1).

민감도는 유족연금의 발생을 event로 보았을 때 유족연금을 지급 받는 사람을 유족연금 수급자로 옳게 예측할 확률을 나타내며, 특이도는 일시금을 지급 받게 되는 것을 non-event로 보았을 때 일시금을 지급받게 되는 사람에 대하여 일시금을 지급 대상자로 옳게 예측하는 확률을 나타낸다. 표 3.4에서 보는바와 같이 threshold= 0.5에서 세 모형을 비교하여 보면 모형 (1)과 (2)는 정확도와 민감도가 모형 (3)보다 높지만 특이도는 각각 21.2%과

표 3.6: 세 모형의 사후확률의 크기에 따른 등급 하한

등급	모델(1)	모델(2)	모델(3)
1	0.97	0.98	0.84
2	0.96	0.97	0.80
3	0.95	0.96	0.76
4	0.93	0.95	0.72
5	0.92	0.93	0.67
6	0.90	0.91	0.61
7	0.88	0.88	0.54
8	0.83	0.80	0.44
9	0.66	0.56	0.22
10	0.10	0.15	0.01

31.6%로 모형 (3)에 비하여 현저하게 낮다. 예컨대, 다음 달에 지급해야 하는 평균 유족 연금액이 30만원이고 동월에 지급해야 하는 평균 일시금 지급액이 100만원이라고 한다면 유족연금을 유족연금으로 예측하지 못했을 때의 위험보다 일시금을 일시금으로 예측하지 못했을 때의 위험이 더 크기 때문에 두 위험의 크기를 고려하여 의사결정을 해야만 한다. 민감도와 특이도의 균형에 있어서 모형 (1)과 (2)는 비대칭을 보이고 있다. 즉, 정확도를 높일 경우 민감도도 높아지는 반면에 특이도가 급격히 떨어지고, 특이도가 높아지면 민감도와 정확도가 떨어진다. 반면, 모형 (3)의 경우는 높은 정확도를 보이는 threshold에서 민감도와 특이도가 균형을 보인다. 모형에 따라 최적의 민감도와 특이도, 정확도를 얻기 위하여, 자료의 사전 확률 비율에 따른 threshold 결정을 표 3.4의 하단에 보였다. 그러나, 모형 (1)과 (2)는 사후 확률의 분포가 지나치게 한 쪽으로 치우친 비대칭을 보이기 때문에 등급별 분류를 위한 threshold의 결정이 어려웠다 (그림 3.1 참조).

합리적인 threshold의 결정은 반응변수의 빈도 비율에 영향을 받는데 수급자들의 등급을 사후 확률의 값에 따라 분류함으로써 연금관리의 효율성을 높이기 위해서는 사후 확률의 합리적 분류를 위한 등급 기준이 필요하다. 가령 예측된 사후확률의 분포가 한쪽으로 치우쳐져 있어서 주어진 기준값(threshold)에 따른 민감도와 특이도가 비대칭의 현상을 보일 경우 등급을 합리적으로 분류하기 위한 기준값을 결정하는 것은 단순하지 않다. 본 연구에서 제안하는 모형 (3)은 다른 모형 (1)과 (2)에 비하여 모형의 Max rescaled R-square가 높고 사후 확률의 분포에 있어서 한 쪽으로 치우친 정도가 많이 줄어들었기 때문에 사후 확률에 따라 등급 기준을 정할 경우 민감도와 특이도의 균형을 맞추는 것이 가능하고, 따라서 유족연금 발생여부의 예측에 있어서 사후확률에 따른 등급별 분류가 용이하므로 분류된 그룹의 특성에 맞는 관리가 가능하다.

표 3.6에서 보는 바와 같이 대상자들의 유족연금수급에 대한 등급별 분류를 위하여 전체 자료의 사후확률을 크기순으로 나열 후 percentile에 따라 등급을 줄 경우 모형 (3)의 등급간격은 모델 (1)과 (2)의 등급간격보다 비교적 일정한 차이를 보임을 확인할 수 있다. 이는 각 모형의 사후 확률 분포의 특성이 반영된 결과로 각 모형에 대하여 대상자들의 연금수급가능성을 사후확률의 크기에 따라 평점할 경우 평점표의 타당성평가를 통하여 그 차

이를 확인할 수 있다.

사후 확률값을 0.1씩 등간격으로 분류했을 때 등급별 평점표의 타당성 평가로 실시한 Kolmogorov-Smirnov 통계량을 그림 3.4에 보였다. K-S통계량은 누적 유족연금지급 비율과 일시금지급 비율 분포의 최대 차이값을 의미하며, 표에서 보는 것처럼 평점이 높아질수록 유족연금지급 비율이 높아지고 일시금지급 비율이 낮아짐을 볼 수 있는데, 이는 대상자들의 연금수급가능성을 잘 순위화하고 있음을 나타낸다. 모형 (3)의 경우 K-S 통계량 값이 평점표의 타당성 평가에서 '매우 우수한 정도'의 여부를 판단하는 기준 값인 50보다 큰 값을 보였으며, 모형 (1), (2)보다 다소 큰 값을 보였다 (강현철 등, 2006, pp. 158). 그림에서 보는 바와 같이 상위등급으로 갈수록 일시금지급의 비율이 감소하고 유족연금의 지급비율이 증가하는 추세가 모형 (1)과 (2)의 경우보다 모형 (3)에서 더 강하게 나타남을 보여 (1), (2)보다는 모형 (3)이 더 잘 순위화 하고 있음을 보여주었다.

예측모형의 성능평가를 위해 세 모형에 대하여 validation 자료에 대한 각 등급의 향상도(리프트)를 그림 3.3에 보였다. 이를 살펴보면, 모형 (3)은 모형 (1), (2)보다 상위등급에서는 더 높은 반응률을, 하위등급에서는 더 낮은 반응률을 보여 모형 (3)이 모형 (1), (2)보다 좋은 성능의 예측 모형임을 확인하였다.

세 모형에 대한 Max rescaled R-square, 정확도와 ROC curve, 평점도의 타당성평가 그리고 리프트 그래프를 통한 모형의 예측력 비교 결과를 근거로 모형 (3)을 유족연금의 발생 예측을 위해 적합한 모형으로 제안하였다. 모형 (2)의 경우는 교호작용효과를 추가함에 따른 복잡한 정도가 커짐으로 인해 모형의 단순성과 주요인의 특성파악이 어려워진 반면 Max rescaled R-square, 정확도, 리프트 그래프를 통한 모형의 예측력 비교에서 향상된 값들을 보여주지 못했다. 모형 (3)에 있어서도, 교호작용을 추가할 경우 등급분류를 위한 주요인의 특성파악은 어려워진 반면 그로 인해 얻는 통계적 특성의 향상이 없었다는 점에서, 대상자관리 방안을 제시하기위한 예측모형에 있어서는 교호작용효과를 고려하지 않는 것이 적합한 모형이라 판단하였다.

제안하는 모형 (3)의 예측식은 다음과 같다.

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -5.28 + 0.05\text{나이} + 0.002\text{기간} - 0.25C1 - 0.07C2 + 0.05\text{소득수준} + 2.68S,$$

여기서 $\hat{p}$은 식(2.1)에서 정의한 확률의 추정치이며, 이는 설명변수가 주어졌을 때 유족연금을 수급할 확률의 추정치이고, $1-\hat{p}$은 설명변수가 주어졌을 때 일시금을 지급할 확률에 대한 추정치이다. $C1$은 가입종별이 사업장 가입일 경우 1, 그렇지 않을 경우 0을 $C2$는 가입종별이 지역가입일 경우 1, 그렇지 않을 경우 0을 나타내는 더미변수이고 S는 성별을 나타내는 더미변수로 남성인 경우는 1, 여성인 경우는 0을 나타낸다.

제안하는 모형을 통해 수급자들에 대하여 예측된 유족연금 수급확률에 따라 등급을 나누어 유족연금 발생에 대해 가능한 관리 방안을 제시하고 등급에 따라 분류된 그룹의 특성을 파악함으로써 유족연금 대상자들의 연금액을 파악할 수 있는 방안을 예시하고자 한다.

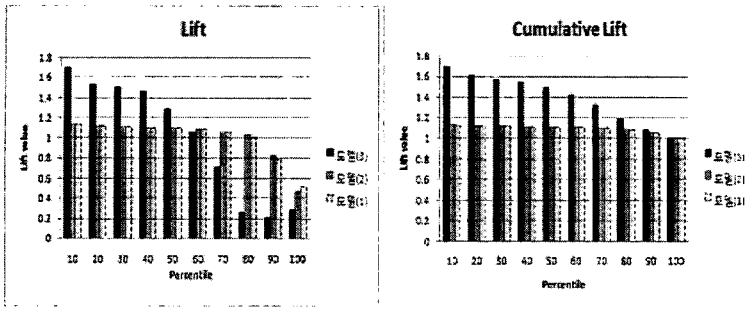
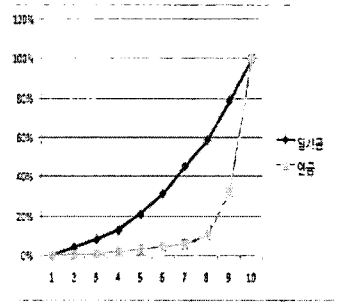
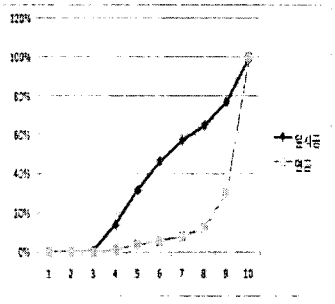


그림 3.3: validation 자료에 대한 리프트 그래프

등급	모델(1) 전체(%) 일시금	전체(%) 연금	누적(%) 일시금	누적(%) 연금	k-s
10	0	0	0	0	0
20	4	0	4	0	4
30	4	1	8	1	7
40	5	1	13	2	11
50	8	1	21	3	18
60	10	1	32	5	27
70	14	1	46	6	40
80	13	4	59	10	48
90	20	22	79	32	47
100	21	68	100	100	0
				K-S	48



등급	모델(2) 전체(%) 일시금	전체(%) 연금	누적(%) 일시금	누적(%) 연금	k-s
10	0	0	0	0	0
20	0	0	0	0	0
30	1	0	1	0	1
40	13	1	14	1	13
50	18	2	32	3	28
60	15	2	47	6	41
70	11	2	57	8	50
80	8	5	65	13	52
90	12	17	77	30	47
100	23	70	100	100	0
				K-S	50



등급	모델(3) 전체(%) 일시금	전체(%) 연금	누적(%) 일시금	누적(%) 연금	k-s
10	19	3	19	3	16
20	18	2	37	5	32
30	14	2	51	7	44
40	9	4	60	11	49
50	9	7	69	18	51
60	10	13	79	32	47
70	8	28	87	52	35
80	9	27	96	78	17
90	4	19	100	97	3
100	0	3	100	100	0
				K-s	51

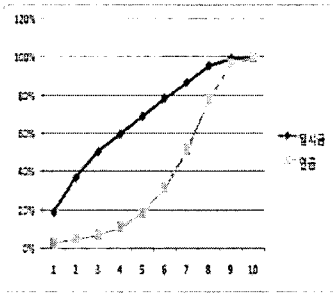


그림 3.4: K-S 통계량 및 유족연금과 일시금지급의 비율 비교

표 4.1: 장애 연금의 경우 추가된 설명 변수

변수	정의	특성
장애등급	장애의 등급을 나타내는 변수	1 = 1등급, 2 = 2등급, 3 = 3등급
장애원인	장애원인을 나타내는 변수	1 = 질병, 2 = 부상

표 4.2: 분석 대상 빈도

유족연금 발생여부	장애 연금의 사망자 전체 빈도수	모형 (3)에 적합한 자료의 빈도수
1 = 유족연금	각 10,616명	7,959 명
0 = 지급종료	각 7,959명	7,959 명

표 4.3: 장애 유족연금의 모수 추정값과 odds ratio

Parameter		Estimate	odds ratio	신뢰 하한	신뢰 상한	Standard Error	Wald Chi-Square	Pr
Intercept		−2.9260	-	-	-	0.1771	272.9749	< .0001
나이		0.0116	1.012	1.007	1.016	0.0022	28.4492	< .0001
소득수준		0.0051	1.005	1.001	1.010	0.0023	5.0723	0.0243
가입기간		−0.0069	0.993	0.992	0.994	0.0004	267.8225	< .0001
장애등급	1 *vs* 3	2.1006	8.171	7.401	9.022	0.0505	1728.089	< .0001
장애등급	2 *vs* 3	1.8074	6.095	5.560	6.681	0.0468	1490.367	< .0001
장애원인	1 *vs* 2	0.4005	1.493	1.325	1.682	0.0609	43.2929	< .0001
성별	1 *vs* 0	1.4441	4.238	3.734	4.810	0.0646	500.3296	< .0001
가입종별	0 *vs* 2	−0.1542	0.857	0.731	1.005	0.0810	3.6236	0.0570
가입종별	1 *vs* 2	−0.4887	0.613	0.530	0.710	0.0749	42.5274	< .0001

4. 기타연금에 대한 예측모형

위의 결과를 토대로, 나머지 두 종류의 연금에 대한 예측모델을 제안하고자 한다. 모델 (3)과 같이 반응변수의 빈도를 동일하게 하여 분석을 실시하였으며, 앞의 분석과 동일한 절차를 따른 결과로 다음의 4.1절과 4.2절에서 소개한다.

4.1. 장애연금

장애연금수급자 가운데 유족연금 발생의 특성을 파악하기 위해 로지스틱 회귀분석을 실시하였다. 앞에서 다룬 '가입자 및 대기자'그룹의 경우 다루었던 다섯 개의 설명변수와 함께 표 4.1의 추가된 변수를 설명변수로 하였다.

로지스틱에서 유족연금수급자들의 특징을 파악하기 위해서 지급종료와 대비하여 분석을 실시하였다. 표 4.3은 로지스틱분석 결과 추정된 모수와 odds ratio를 보여주고 있다.

장애연금에 대한 유족연금의 발생에 영향을 미치는 요인들을 살펴보면 연속형 변수인 경우, 사망연령이 높을수록, 소득수준이 증가할수록, 가입기간이 짧을수록 유족연금이 발

표 4.4: 분석 대상 빈도

유족연금 발생여부	조기노령연금의 사망자 전체 빈도수	모형 (3)에 적합한 자료의 빈도수
1 = 유족연금	각 68,640 명	20,536 명
0 = 지급종료	각 20,536 명	20,536 명

표 4.5: 노령 유족연금의 분석결과 추정된 모수와 odds ratio

Parameter		Estimate	odds ratio	신뢰 하한	신뢰 상한	Standard Error	Wald Chi-Square	Pr
Intercept		−1.8615	-	-	-	0.2212	70.8212	< .0001
연령		−0.0079	0.992	0.986	0.998	0.00309	6.6052	0.0102
가입기간		0.0030	1.003	1.002	1.004	0.000489	36.2295	< .0001
소득수준		0.0123	1.012	1.010	1.015	0.00121	103.1076	< .0001
가입종별	0 *vs* 2	−0.1774	0.837	0.781	0.898	0.0354	25.0924	< .0001
가입종별	1 *vs* 2	−0.0453	0.956	0.909	1.005	0.0256	3.1353	0.0766
성별	1 *vs* 0	2.4493	11.58	10.822	12.391	0.0345	5025.8060	< .0001

생할 가능성이 높았다. 또한 범주형의 변수를 해석해 보면, 장애등급이 3등급에 비해 1등급 또는 2등급일 경우, 장애원인이 부상에 비해 질병일 경우, 성별이 여성에 비해 남성일 경우, 유족연금이 발생할 가능성이 높았으며, 가입종별이 임의 가입에 비해 지역가입 또는 사업장가입일 경우 유족연금이 발생할 가능성이 낮았다.

장애 연금수급자에 대하여 유족연금 발생에 대한 모형 (3)의 예측식은 다음과 같다.

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -2.926 + 0.0116\text{나이} + 0.00506\text{소득수준} - 0.00692\text{가입기간} + 2.1006D1 + 1.807D2 + 0.4005\text{장애원인} + 1.1441S - 0.1542C1 - 0.4887C2,$$

여기서 $\hat{p}$은 식 (2.1)에서 정의한 확률의 추정치이며, 이는 설명변수가 주어졌을 때 유족연금을 수급할 확률의 추정치이고, $1-\hat{p}$은 설명변수가 주어졌을 때 장애연금의 지급이 종료될 확률에 대한 추정치이다. 장애원인은 질병일 경우 1, 부상일 경우를 0으로, $D1$은 1급 장애를 1로 그렇지 않으면 0으로, $D2$는 2급 장애는 1로 그 외에는 0인 더미변수이다. $C1$은 가입종별이 사업장 가입일 경우 1, 그렇지 않을 경우 0을 $C2$는 가입종별이 지역가입일 경우 1, 그렇지 않을 경우 0을 나타내며, S는 남성인 경우를 1로 여성인 경우는 0을 나타내는 더미변수이다.

4.2. 조기노령연금

조기노령연금수급자 가운데 유족 연금으로 전환된 경우의 특징을 파악하기 위해 로지스틱 회귀분석을 실시하였다. 여기서 사용된 변수에 대한 정보는 표 4.4과 같다.

로지스틱에서 유족연금수급자들의 특징을 파악하기 위해서 지급종료의 경우와 대비하여 분석을 실시하였다. 표 4.5은 로지스틱분석 결과 추정된 모수와 odds ratio를 보여주고

표 5.1: 예시된 등급별 특성

사후확률	등급	특성
0.75–1.00	1	평균 가입기간이 70 소득수준 39 사망나이 53 성별이 남성 가입종별이 사업장 가입
0.50–0.75	2	평균 가입기간 48 소득수준 15 사망나이 48 성별이 남성 지역가입
0.25–0.50	3	평균 가입기간 32 소득수둔 12 사망나이 34 성별이 남성 지역가입
0.00–0.25	4	평균 가입기간 37 소득수준 12 사망나이 43 성별이 여성 지역가입

있다.

조기노령연금에 대한 유족연금의 발생에 영향을 미치는 요인들을 살펴보면 연속형 변수인 경우, 사망연령이 낮을수록, 소득수준이 증가할수록, 가입기간이 길수록 유족연금이 발생할 가능성이 높았다. 또한 범주형의 변수를 해석해 보면, 가입종별이 임의가입 대비 사업장가입 또는 지역가입일 경우 유족연금일 가능성이 낮았으며, 성별이 여성에 비해 남성일 경우 유족연금을 받을 가능성이 높았다.

조기노령연금 수급자에 대하여 유족연금 발생에 대한 모형 (3)의 예측식은 다음과 같다.

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -1.8615 - 0.00794\text{나이} + 0.00295\text{가입기간} + 0.0123\text{소득수준} - 0.1774C1 \\ -0.0453C2 + 2.4493S,$$

여기서 $\hat{p}$은 식 (2.1)에서 정의한 확률의 추정치이며, 이는 설명변수가 주어졌을 때 유족연금을 수급할 확률의 추정치이고, $1-\hat{p}$은 설명변수가 주어졌을 때 조기노령연금의 지급이 종료될 확률에 대한 추정치이다. $C1$은 가입종별이 사업장 가입일 경우 1, 그렇지 않을 경우 0을 $C2$는 가입종별이 지역가입일 경우 1, 그렇지 않을 경우 0을 나타내며, S는 남성인 경우를 1로 여성인 경우는 0을 나타내는 더미변수이다.

5. 유족연금 관리방안 예시

효율적인 유족연금관리를 위한 방안의 예시로 '가입자 및 대기자'그룹의 경우 유족연금 수급대상자들의 등급을 네 등급으로 나누는 경우를 표 5.1에 예시하였다. 모형 (3)에 적용하였을 때, 예측된 사후 확률분포를 0.25씩 4등급으로 나누고 각 등급에 속한 대상자들의 특성을 보였다.

각 대상자들의 예측된 사후 확률값이 0.5보다 클 경우 유족연금 지급을 받을 것으로, 0.5보다 작을 경우 일시금 지급을 받을 것으로 분류한다. 예측된 사후 확률분포를 0.25씩 등간격으로 분류해 보면 1등급의 경우 가입기간이 길고 소득수준이 높아 매월 수급되는 연금액이 가장 많을 것으로 예측되는 등급이며 유족 연금 수급 확률이 0.75 이상인 그룹으로 유족 연금 수급 가능성이 가장 높은 그룹이어서 장기적인 관리가 필요한 등급이라 할 수 있으며, 4등급의 경우 사망일시금이 지급될 가능성이 가장 높은 등급으로 유족 연금 수급 확률은 0.25 이내인 그룹이다. 일시금으로 지급 될 확률이 상대적으로 더 높기 때문에 장기보다는 단기적인 관리가 필요한 대상이다. 이와 같은 방법으로 각 등급에 속하는 대상자들

을 예측하고 그에 따른 유족연금액 및 일시금을 산정하는 과정에서 등급의 특성에 맞는 적절한 가중치를 줌으로써 예상 유족연금액을 예측한다면 예측오차를 줄여 합리적인 예측이 가능할 것으로 기대한다.

장애연금과 조기노령연금 수급자그룹의 경우에도 위에서 제시한 바와 같은 과정을 통하여 예측된 사후 확률 값을 이용한 등급에 따라 유족연금 관리 방안을 제안할 수 있다.

6. 요약 및 결론

2007년 7월 23일 국민연금법이 개정되면서 국민연금에 대한 관심이 높아지고 있지만 저출산, 고령화 사회가 초래할 수 있는 연금재정 적자현상에 대한 뚜렷한 방안이 제시되고 있지 못하고 있다 (양준모, 2006).

본 연구에서는 효율적인 연금관리를 위하여 1988년 1월부터 2007년 5월까지의 국민연금 자료에서 유족연금을 중심으로 그 발생에 대한 요인을 찾고 이를 예측하기 위하여 두 단계의 로지스틱 회귀분석을 실시하였다.

첫 번째의 로지스틱 분석에서는 전체 대상자에 대해 유족 연금 발생에 영향을 미치는 요인의 특성을 파악하기 위하여 세 그룹의 전체 사망자를 대상으로 분석을 실시하였으며, 두 번째 로지스틱 분석에서는 '장애연금'이나 '조기노령연금' 수급자 그룹에 비해 가장 높은 확률의 유족 연금 발생을 보이는 '가입자 및 대기자'그룹의 사망자를 대상으로 유족연금의 발생에 관한 통계적 모델을 고려하였다. 합리적이고 적합도가 높은 모델을 제안하기 위하여 반응변수의 빈도수를 동일하게 맞춘 자료에 적합한 모델을 고려하였고 이를 일반적인 로지스틱 모델과 비교하였다. 제안한 모델의 경우 threshold = 0.5에서 전체 자료에 대한 정확도는 80.78%를 보였으며, 최적의 threshold를 선택할 경우 모델 (1)과 (2)의 정확도는 각각 80.70%와 78.90%를 보였다. 각 모형에 있어서 최적의 threshold를 고려할 경우 미소한 차이이지만, 민감도는 제안한 모형 (3)에서 특이도는 모형 (2)에서 가장 높았다. 그러나 세 모형 모두 동일한 ROC curve를 보임으로, 특정 threshold에 상관없이 전반적인 정확도는 세 모형 모두 동일함을 보였다.

모델 (3)의 경우는 사후 확률의 분포가 비교적 완만한 반면 모델 (1)과 (2)의 경우는 사후 확률 분포의 꼬리가 지나치게 왼쪽으로 치우쳐 있어서 사후확률에 따른 등급별 분류가 용이하지 않았다.

예측 검정력을 위한 Max-rescaled R-square값에서도 모형 (3)은 모형 (1)과 (2)의 경우보다 각각 35.8%와 18% 증가한 값을 보였다. 이에 따라 민감도와 특이도 그리고 사후확률의 분포를 동시에 고려할 경우, 변별력이 높은 등급분류 기준을 찾는데 있어서 합리적인 모형으로 모델 (3)을 제안하였다. 이는, 모델 (3)의 경우 유족연금수급 대상자의 등급별 분류가 용이하여서 이에 따라 유족연금수급자의 등급별 관리뿐만 아니라 연금재정 관리에 있어서도 유족 연금액 및 일시금 지급액에 대한 현실적 예측이 가능하기 때문이다.

유족연금을 효율적으로 관리하기 위하여 수급자들을 등급에 따라 분류하여 관리하는 방안을 제안하였는데, 유족연금 수급에 대한 사후확률에 따라 수급자들을 4등급으로 분류할 경우, 등급별 특성을 살펴보면 등급에 따른 소득수준의 차이, 성별, 예상되는 연금 지급

액의 차이, 유족 연금의 형태로 지급받을 확률 등의 파악이 가능하고 이에 따른 등급별 관리 방안이 가능함을 예시하였다.

연금을 운영함에 있어 각 그룹에 속하는 대상자들의 특성을 고려한다면 그에 따른 예상 연금액을 보다 효율적으로 산출할 수 있어 연금지급액을 바르게 예측하지 못하여 발생하는 문제와 비효율적인 연금운용으로 인하여 발생할 수 있는 기회비용을 줄임으로써 부적절한 연금 운용으로 인한 문제를 줄일 수 있다고 본다.

본 연구에서는 연금을 관리하는데 사용되는 기본적인 요인들을 설명변수로 하여 각 설명변수가 유족연금의 발생에 영향을 미치는 정도와 대상자들의 유족연금발생 가능성에 대한 확률을 예측하였다. 연금예측에 영향을 미치는 추가적인 변수에 대한 정보와 이에 대한 자료가 있다면 더 높은 정확도를 보이는, 효율적인 모형을 구축할 수 있고 이에 따라 추정된 확률 값도 더 정확하게 산출할 수 있다. 주어진 변수 외에 가입자들의 특성을 파악할 수 있는 다른 요인들을 이용한 분석은 제한된 데이터의 특성상 고려하기 어려웠다. 그러므로 제시된 등급에 따른 이러한 요인들을 활용하여 연금수급자들의 특징들을 파악하고 이를 관리하는 방안에 대한 추가적인 연구가 필요하다.

참고문헌

강현철, 한상태, 최종후, 이성건, 김은석, 엄익현, 김미경 (2006). <고객관계관리 (CRM)를 위한 데이터마이닝 방법론>, 자유아카데미, 경기.

김지훈 (2000). 공무원연금 재정전망, <응용통계연구>, **13**, 19–34.

안홍순 (2007). 국민연금 재정안정화를 위한 기본연금산식의 구조개혁, <사회보장연구>, **23**, 297–326.

양준모 (2006). 고령화와 국민연금: 고령화에 따른 국민연금제도 개혁방향의 모색, <한국경제의 분석>, **12**, 113–182.

이근홍 (2006). 국민연금의 개혁과 경로연금의 과제, <노인복지연구>, **32**, 7–29.

Bergh, H., Baigi, A., Månsson, J., Mattsson, B. and Marklund, B. (2007). Predictive factors for long-term sick leave and disability pension among frequent and normal attenders in primary health care over 5 years, *Public Health*, **121**, 25–33.

Huberman, G., Iyengar, S. S and Jiang, W. (2007). Defined contribution pension plans: Determinants of participation and contributions rates, *Journal of financial Services Research*, **31**, 1–32.

Humphreys, M., Costanzo, P., Haynie, K. L, Ostbye, T., Boly, I., Belsky, D. and Sloan, F. (2007). Racial disparities in diabetes a century ago: Evidence from the pension files of US Civil War veterans, *Social Science and Medicine*, **64**, 1766–1775.

SAS (2000). SAS System for Windows V.9.1, SAS Institute Inc.

[2007년 11월 접수, 2008년 1월 채택]

Analysis on the Survivor's Pension Payment with Logistic Regression Model*

Mijung Kim[1] Jin Hyung Kim[2]

ABSTRACT

Research for efficient management of the National Pension has been emphasized as the current society trends toward aging and low birth rate. In this article, we suggest a statistical model for effective classification and prediction of the reserve for the survivor's pension in Korea. Logistic regression model is incorporated; correct classification rate, and distribution of the posterior probability for the reserve of survivor's pension are investigated and compared with the results from the general logistic models. Assessment of predictive model is also done with lift graph, ROC curve and K-S statistic. We suggest strategies for reducing financial risks in managing and planning the pension as an application of the suggested model.

Keywords: The national pension, survivor's pension, predictive model assessment, logistic regression model.

* Data are from National Pension Service in Korea.

1) Corresponding author. Research Professor, Institute for Mathematical Sciences, Yonsei University, 134 Shinchon-Dong, Seodaemun-Gu, Seoul 120-752, Korea.
E-mail: mjkim@yonsei.ac.kr

2) Graduate Student, Dept. of Industrial & Information Engineering, Yonsei University, 134 Shinchon-Dong, Seodaemun-Gu, Seoul 120-752, Korea.
E-mail: statkjh@yonsei.ac.kr