Annals of Statistical Research
Vol. 6. No. 1, 2009, pp.1~25

Panel Data Analysis : A study on quality of life in low income

Chongwon Chang[1], Eunyoung Kim[2], Hyunjung Noh[3]

Abstract

Panel data is observed and collected repeatedly over time from the same sample. Therefore panel data is in the form of cross section data combined with time series data, and usually it is constructed from continuous survey so find individuals, households or companies.

Other countries have conducted panel surveys since the mid-1900s. On the other hand, in Korea it has been started since 1990 and maintained mostly by research institutions. The collected panel data is supplied every year to researchers who want to analyze it. However, most published papers have dealt with a one-year data set and only a few papers used the multi-year panel data, probably because proper statistical software are hard to find.

The purpose of this study is to investigate the theory of the panel data analysis and is to analyze a multi-year data from the Korea Welfare Panel Study(KWPS).

Key words : Panel data analysis, Low income, Quality of life, Korea Welfare Panel Study(KWPS), Life satisfaction

1) Doctoral Candidate for Statistics, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : jjangkid@dongguk.edu
2) Graduate student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : key0223@dongguk.edu
3) Graduate student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : hjhj2143@dongguk.edu

제1장 서 론

패널자료(panel data)는 동일 표본으로부터 여러 시점에 걸쳐서 반복적으로 수집한 자료이다. 따라서 패널자료는 일정한 시점에 표본들이 어떤 값을 갖는지를 보여주는 횡단면자료(cross-section data)에 표본들이 시간의 흐름에 따라 어떤 값을 갖는지에 대한 변화를 보여주는 시계열자료(time-series data)의 형태가 추가된 자료라고 볼 수 있으며, 주로 개인이나 가구, 기업들을 표본으로 하여 일정 주기마다 지속적으로 조사한다.

횡단면적인 정보뿐만 아니라 시계열 정보까지 보유하고 있는 패널자료는 시계열 분석 내지 횡단면 분석만으로 파악할 수 없는 추가적 정보를 얻을 수 있는 정보의 보고라고 할 수 있다. 무엇보다 실증분석에 있어서 패널자료분석이 가장 효과적인 방법이 될 수 있음은 패널자료만이 가지는 장점이 있기 때문인데 Hsiao(1985), Klevmarken(1989), Solon(1989)는 패널자료분석에 대해 다음과 같은 장점을 들고 있다. 먼저 패널자료는 개별적 특이성을 통제할 수 있으며, 연구자에게 다양한 정보를 제공해주고 다중공선성의 문제를 줄일 수 있다는 것이다. 다음으로 조정의 동태성을 가능하게 해주며 순수한 횡단면이나 순수한 시계열 데이터에서 포착하기 힘든 효과를 보다 잘 측정할 수 있다는 것이다. 마지막으로 패널자료는 횡단면자료나 시계열자료에 비해서 복잡한 행태적 모형을 구축 및 검증할 수 있으며 개인, 기업, 정부 등과 같이 미시적인 단위에서 수집되는 자료에서 발생하는 편이(bias)를 통제해준다는 것이다.

외국에서는 1900년대 중반 이후부터 다양한 패널자료를 구축해오고 있다. 해외 주요 국가들의 대표적인 패널조사에는 미국의 PSID(Panel Study of Income Dynamics), HRS(Health and Retirement Study), NLS(National Longitudinal Survey)와 독일의 GSOEP(German Socio Economic Panel), 영국의 BHPS(British Household Panel Study), ECHP(European Community Household Panel Study),그리고 캐나다의 SLID(Survey of Labor Income Dynamics)등이 있다.

우리나라는 대우경제연구소에서 1993년부터 1998년까지 총 6년간 실시한 한국가구패널조사(KHPS : Korean Household Panel Study)가 최초의 패널조사이다. 1998년부터 한국노동연구원에서 조사·발표하고 있는 한국노동패널(KLIPS : Korea Laborand Income Panel Study)은 11차 조사(2008년)까지 완료되었다. 그 외에 한국고용정보원의 청년패널조사(Youth Panel), 한국직업능력개발원의 인적자본기업패널(HCCP : Human Capital Corporate Panel Survey)등 국내의 패널조사는 연구기관들을 중심으로 다양한 분야에서 이루어지고 있다.

이 자료들은 많은 연구자들에 의해 분석되어 학술대회 등을 통해 논문이 발표되고 있다. 그러나 패널자료임에도 불구하고 대개 한해년도 자료만 가지고 분석하거나 다년도 자료를 사용하더라도 개별연도의 분석결과를 비교하는 방법을 사용하고 있다. 이처럼 다년도 자료의 패널분석이 적은 것은 다년도 조사자료 처리의 어려움 등의 이유도 있겠지만 그 중에서도 패널자료분석을 위한 적당한 소프트웨어가 제공되지 않았기 때문에 다년도 자료의 분석이 어려웠을 것으로 짐작된다.

현재 패널자료분석을 목적으로 개발된 도구로는 SAS/ETS의 PROC TSCSREG와 SAS/STAT의 PROC MIXED 그리고 R의 nlme 패키지 등이 있으며, 그 외에도 SPSS, STATA, LIMDEP 등이 패널자료분석에 활용되고 있다.

본 논문에서는 한국보건사회연구원의 주관으로 조사된 한국복지패널(KWPS : Korea Welfare Panel Study)자료를 SAS/STAT의 PROC MIXED를 사용하여 분석하고자 한다. 한국복지패널은 빈곤층, 근로빈곤층(working poor), 차상위층(near poor)의 규모와 상태변화를 동적으로 파악하여 정책지원을 위한 기초자료를 생산하고, 소득계층별, 경제활동 상태별, 연령별 등 각 인구집단의 생활실태와 복지욕구 등을 역동적으로 파악하여 정책의 효과를 평가함으로써 정책형성과 피드백에 기여하고자 하는 목적으로 2006년도부터 약 7,000가구의 표본규모로 조사되고 있다.

한국복지패널의 조사대상 중 저소득가구(low-income)는 표본의 50%를 할당할 정도로 중요한 관심사이며, 1960년대 이후 급격한 경제성장에서 발생하는 사회적 관심사 중의 하나이다. 이에 본 논문에서는 저소득가구의 삶의 질에 대해 연구하고자 하였으며, 이를 위해 저소득가구의 삶의 질에 영향을 미치는 요인들이 무엇인지에 대해 살펴보았다. 또한 일반가구의 삶의 질에 영향을 미치는 요인들과 비교함으로써 저소득가구와 일반가구 간의 차이를 제시하고자 하였다.

한국보건사회연구원에서의 '한국인의 삶의 질과 과제'(1997)에서는 삶의 질이란 인간생활의 질적 수준을 나타내는 의미로 개인의 주관적 만족도 또는 행복감을 뜻하는 포괄적인 의미를 갖는 것으로 정의하였다. 그리고 삶의 질과 유사한 개념으로는 웰빙, 복지, 생활만족도, 행복감, 사회지표, 생활수준, 주관적 웰빙, 생활의 질 등을 표현하였다. 이 중 생활만족도는 한국복지패널에서 조사된 문항 중 하나로써 본 논문에서 삶의 질을 평가하는 척도로 간주하여 패널자료분석에 사용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 패널자료의 개념 및 특징, 국내외 패널조사현황 그리고 패널자료의 구조와 패널모형의 원리 등에 대해 알아본다. 3장에서는 한국복지패널 조사의 개요를 살펴보고 이를 이용하여 저소득가구의 생활만족도에 영향을 미치는 요인

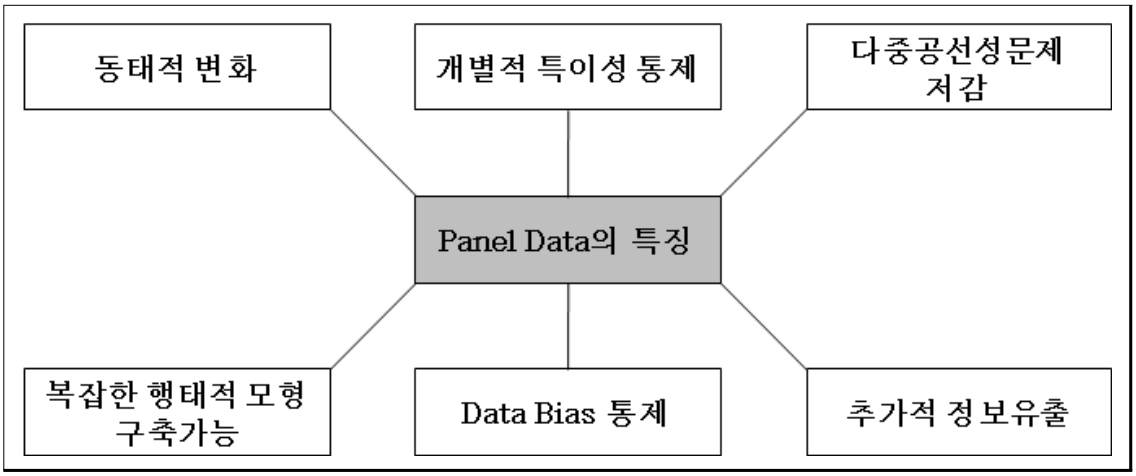
에 대해 알아본다. 또한 생활만족도에 영향을 미치는 요인에서 저소득가구와 일반가구 간의 차이를 비교한다. 마지막으로 4장에서는 본 연구의 결과를 정리하고 향후과제를 제시한다.

제2장 패널모형(Panel model)

2.1 패널자료의 개념

패널모형이란 패널자료를 이용한 분석으로서 시계열 분석과 횡단면 분석을 동시에 수행하는 분석모형을 의미한다. 패널자료는 동일 표본으로부터 여러 시점에 걸쳐서 반복적으로 수집한 자료로서, 횡단면적인 데이터 정보뿐만 아니라 시계열 데이터 정보를 보유하고 있어 시계열 분석 내지 횡단면 분석만으로 파악할 수 없는 추가적 정보를 얻을 수 있는 정보의 보고라고 할 수 있다. 무엇보다 실증분석(empirical research)에 있어서 패널분석이 가장 효과적인 방법이 될 수 있음은 패널자료만이 가지는 여러 가지 장점이 있기 때문인데 Hsiao(1985), Klevmarken(1989), Solon(1989)는 패널자료 분석에 다음과 같은 장점을 들고 있다(Baltagi, 2005).

첫째, 패널자료는 개별적 특이성(individual heterogeneity)을 통제할 수 있다는 것이다. 개별적인 특이성을 통제하지 못할 경우 시계열 분석이나 횡단면 분석은 왜곡된 결과를 얻을 위험이 커지게 되는데, 패널자료 분석은 시계열 분석이나 횡단면 분석에서는 불가능한 개별특성효과(individual effect)와 시간특성효과(time effect)를 모두 통제할 수 있는 장점이 있다. 둘째, 패널자료는 연구자에게 다양한 정보를 제공해주며 다중공선성의 문제를 줄일 수 있다는 것이다. 게다가 보다 많은 자유도(degrees of freedom)와 가변성(variability)을 제공해주어 분석을 용이하게 해준다. 셋째, 패널자료는 조정의 동태성(dynamics of adjustment)을 가능하게 해준다는 것이다. 상대적으로 안정된 횡단면 분포에서는 포착하기 힘든 다양한 변화를 포착하게 도와준다. 넷째, 패널자료는 순수한 횡단면이나 순수한 시계열자료에서 포착하기 힘든 효과를 보다 잘 측정해낼 수 있다는 것이다. 다섯째, 패널자료는 횡단면 자료나 시계열 자료에 비해서 복잡한 행태적 모형을 구축 및 검증하게 해준다. Hsiao(1986)는 시차모형(lag model)에 있어서도 패널자료가 시계열자료보다 자료에 대한제약이 덜 가해지기 때문에 효과적이라고 하였다. 여섯째, 패널자료는 개인, 기업, 정부 등과 같이 미시적인 단위에서 수집되는 데이터에서 발생하는 편이(bias)를 통제하게 해준다. 이것은 두 번째의 개별특성효과와 비슷한 것으로 개별 데이터집합에서 생길 수 있는 각 종의 편이들을 제거하여 분석할 수 있음을 의미한다.



<그림 2-1> 패널자료의 특징

유사하게 이영훈(2001)과 전승훈 외(2004)는 횡단면자료와 시계열자료에 비해 패널자료가 갖는 장점을 다음과 같이 정리하였다. 첫째, 표본의 크기가 커지기 때문에 자유도가 늘어남에 따라 추정의 효율성(efficiency)이 향상된다. 둘째, 설명변수 간의 다중공선성(multicollinearity) 문제가 발생할 가능성이 적어진다. 셋째, 추정량의 편의(bias)를 감소시킨다. 넷째, 개인의 관측되지 않는 특성을 반영하여 생략된 변수문제를 해결해준다.

또한, 패널모형은 시계열과정에서 발생하는 추정오차와 지역별 단위의 자료에서 발생하는 추정오차를 통제할 수 있는 장점을 가지고 있기 때문에 횡단면 또는 시계열자료에 비해 현실을 보다 제대로 분석할 수 있는 장점이 있다(Baltagi, 2005). 일반적으로 회귀방정식을 설정할 때 종속변수에 영향을 미치는 모든 변수를 포함할 수는 없다. 설사 모든 변수를 포함시킨다고 하더라도 그것이 가장 좋은 모형이라고 판단하기도 어렵다. 하지만 중요한 것은 종속변수에 매우 중요한 영향을 미침에도 불구하고 독립변수로 포함되지 않은 요인들이 있을 경우 추정된 모형이 매우 위험하게 된다. 패널모형은 이러한 누락된 변수(omitted variable)에 대한 한계를 극복하는 데에 가장 큰 의의를 가지고 있다.

2.2 국내외 패널조사 현황

외국에서는 미국, 독일, 영국 등 선진국들을 중심으로 패널조사가 진행 또는 완료되었다. 대표적인 패널조사로는 미국 미시간 대학 사회조사연구소(Institute for Social Research at the University of Michigan)에서 1968년부터 현재까지 조사 중인 PSID(Panel Study of

Income Dynamics)와 미국 노동통계국(Bureau of Labor Statistics)에서 1966년부터 현재까지 조사 중인 NLS(National Longitudinal Surveys)가 있다. Free(2004)와 Baltagi(2005)는 각 나라별 패널조사 사례들에 대하여 조사기관과 시기, 대상, 표본수 등을 정리하였다. <표 2-1>은 해외의 주요 패널조사 현황을 정리한 것이다.

<표 2-1> 해외 주요 패널조사 현황

구 분	국 가	조사연도	조사주기	표본단위	표본수
PSID	미 국	1968~	2년	가구	6,434가구
NLSY97	미 국	1997~	1년	개인	9,000명
GSOEP	독 일	1984~	1년	가구, 개인	6,600가구, 12,700명
BHPS	영 국	1991~	1년	가구	5,000가구
SLID	캐나다	1993~2000	1년	가구	37,000가구
JPSC	일 본	1994~2000	1년	개인(여성)	2,000명

출처 : 안두진(2008), “SAS PROC PANEL을 이용한 패널자료분석”, 숭실대학교 석사학위논문

우리나라는 대우경제연구소의 한국가구패널조사(KHPS : Korean Household Panel Study)가 1993년부터 1998년까지 6년간 조사가 완료되었고, 한국노동연구원의 한국노동패널(KLIPS : Korea Labor and Income Panel Study)은 1998년에 시작되어 2008년 현재 11차 조사가 진행 중이다. 한국노동패널이 시작된 이후 2000년대 초반부터 국책연구기관들을 중심으로 다양한 주제로 패널조사가 진행되고 있다.

<표 2-2> 국내 주요패널조사 현황(2008년 기준)

구 분	조사기관	조사연도	조사주기	표본단위	표본수
한국가구패널	대우경제연구소	1993 ~1998	1년	가구 개인	4,547가구 (12,032명)
한국노동패널	한국노동연구원	1998~	1년	가구 가구원	5,000가구 (13,000명)
사업체패널		2006~	2년	사업체	2,000개
고령화연구패널		2006~	2년	개인	10,000명
청년패널	한국고용정보원	2001~	1년	개인	8,296명
한국청소년패널	한국청소년정책연구원	2003 ~2008	1년	개인 학부모	2,949명 ~3,697명
한국교육고용패널	한국직업능력개발원	2004~	1년	개인 가구,학교	6,000명
인적자본기업패널		2005~	2년	기업	450개
한국복지패널	한국보건사회연구원	2006~	1년	가구 가구원	7,000가구
여성가족패널	한국여성정책연구원	2007~	1년	개인 (여성)	10,000명
장애인고용패널	노동부, 한국장애인고용촉진공단	2008~	1년	개인 (장애인)	5,000명
한국의료패널	국민건강보험공단, 한국보건사회연구원	2008~	1년 ~2년	가구 가구원	8,000가구 (20,000명)

출처 : 안두진(2008), “SAS PROC PANEL을 이용한 패널자료분석”, 숭실대학교 석사학위논문

2.3 패널자료구조

패널자료란 N개의 관측대상을 T기간 동안 조사한 자료로서 패널자료의 구조는 <표 2-3>과 같다.

<표 2-3>에서 Y는 반응변수이고 $X_1, X_2, \ldots, X_K$는 설명변수이다. $y_{it}(i=1, 2, \ldots, N;$ $t=1, 2, \ldots, T)$는 i번째 관측대상의 t번째 시점에 대한 반응변수의 관측값이고, x_{kit}는 i번째 관측대상의 t시점에 대한 k번째 설명변수의 관측값을 의미한다.

<표 2-3> 패널자료의 구조

CS	TS	Y	X_1	X_2	⋯	X_K
1	1	y_{11}	x_{111}	x_{211}	⋯	x_{K11}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	T	y_{1T}	x_{11T}	x_{21T}	⋯	x_{K1T}
2	1	y_{21}	x_{121}	x_{221}	⋯	x_{K21}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	T	y_{2T}	x_{12T}	x_{22T}	⋯	x_{K2T}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	1	y_{N1}	x_{1N1}	x_{2N1}	⋯	x_{KN1}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	T	y_{NT}	x_{1NT}	x_{2NT}	⋯	x_{KNT}

※ CS(Cross Section), TS(Time Series)

패널자료는 각 관측대상의 관측횟수에 따라 균형(balanced)패널과 불균형(unbalanced)패널로 구분된다. i번째 관측대상의 관측횟수를 $n_i(i=1, 2, \ldots, N)$라 했을 때, 모든 i에 대하여 $n_i = T$를 만족하면 균형패널이고 만족하지 못하면 불균형패널이다. 임의의 관측대상에 대해서 특정시점에서 모형에 사용되는 하나 이상의 변수가 결측값인 경우 해당 시점은 모두 결측값으로 처리가 된다. 예를 들어 첫 번째 관측대상의 첫 번째 관측시점에서 첫 번째 설명변수, 즉 x_{111}이 결측값이면 모형에 사용되는 다른 변수들$(y_{11}, x_{211}, \ldots, x_{K11})$도 모두 결측값으로 처리가 되어 $n_1 = T-1$이 되고 다른 관측대상들과 관측횟수가 다르므로 불균형패널이 된다.

2.4 패널모형의 원리

패널모형은 누락된 변수를 제어하기 위해서 오차항을 개인(individual)간에는 다르나 시간 변동이 없는 변수, 시간변화에 따라 변동하나 개인 간에는 차이가 없는 변수, 개인 간에도 차이가 있고 시간변화에 따라서도 변동하는 확률적 교란항으로 구분하여 다루게 된다. 이를 일반적인 선형모형으로 표현하면 다음 식과 같다(Ashenfelter, 2006).

$$Y_{it} = \alpha + X_{kit}\beta + \epsilon_{it} \quad (\text{ 단, } \epsilon_{it} = \mu_i + \lambda_t + \upsilon_{it}, \quad i = 1, 2, \ldots, N, \; t = 1, 2, \ldots, T)$$

μ_i = 관찰되지 않은 개별특성 효과(unobservable individual effect)

λ_t = 관찰되지 않은 시간 효과(unobservable time effect)

υ_{it} = 확률적 교란항(remainder stochastic disturbance term)

이러한 패널모형은 오차항의 고려방식에 따라 일원오차성분회귀모형(one-way error component regression model)과 이원오차성분회귀모형(two-way error component regression model)으로 나누어지며 오차항에 대한 가정에 따라서 고정효과모형(fixed effect model)과 확률효과모형(random effect model)로 나뉜다.

일원오차성분회귀모형은 시간의 흐름에 따라 변하지 않으며 관찰되지 않은 특정한 변수가 개인마다 잠재해 있다는 가정을 하는 고정효과모형과 시간에 따라 변한다고 가정하는 확률효과모형으로 나뉜다. 반면 이원오차성분회귀모형은 시간의 흐름에 따라 변하지 않고 관찰되지 않는 특정한 변수가 개인마다 잠재해 있고 시계열별 독특한 특성이 각 개인에 잠재해 있음을 가정하는 고정효과모형과 개인마다 시간마다 모두 고정되지 않고 확률적으로 변화한다고 가정하는 확률효과모형으로 나뉜다.

<표 2-4> 패널모형의 구분

구 분	Fixed effect Model	Random Effect Model
One-way error component regression model	Ⅰ	Ⅱ
Two-way error component regression model	Ⅲ	Ⅳ

<표 2-4>의 모형 중 하나를 선택하는 방법은 학자들마다 다르게 나타나지만 일반적으로는 다음과 같이 설명할 수 있다.

모형을 선택하는 가장 좋은 방법은 시간불변의 개별특성효과가 독립변수들과 관련이 되어 있는가를 살펴보는 것이다. 관련이 있게 되면 고정효과모형을 쓰게 되며 관련이 없을 경우 확률효과모형을 선택하게 된다(Johnston, 1997). 흥미로운 것은 확률효과모형이 유효한 경우라도 고정효과모형에 의해 산출된 계수 값은 여전히 일치추정량(consistent estimates)을 제공한다는 것이다. 그 때문에 연구자들은 시간불변의 특정요소가 독립변수들과 관련되어 있는지에 대한 확실한 정보가 없을 경우 확률효과모형보다 고정효과모형을 선호하는 경향이 있다(Johnston, 1997).

고정효과모형과 확률효과모형 어느 쪽도 완벽한 모형은 되지 못한다(Johnston, 1997). 하지만 실증분석에 있어서 두 모형에 대한 합리적인 선택을 하기 위해서는 장단점을 살펴보는 것도 의미가 있다. 고정효과모형의 가장 큰 장점은 개인마다 개별특성효과를 구분하여 계수를 추정한다는 데에 있다. 하지만 개별특성효과를 반영하는 더미변수를 생성하는 과정에서 너무 많은 자유도를 소모하게 되어 결과적으로 독립변수들에 대한 계수값 추정이 상대적으로 정확성을 잃게 된다는 것이다. 확률효과모형의 경우 고정효과모형처럼 더미변수를 설 잃게 과정에서 계수값 추정에 정확성이 떨어지게 되게 위험이 적지만 다소 엄격한 을 이 따르는 것이 흠이다. 왜냐하면 개별특성효과가 독립변수와 전혀 관계를 가질 수 없으며, 실제 분석에서 이를 충족시키기란 여간 어려운 일이 아니기 때문이다(Ashenfelter, 2006). 그렇게 되면 두 모형을 선택하게 해주는 어떠한 기준이 필요하게 된다. 실증분석에 있어서 두 모형 중에 어느 것이 더욱 적합한지에 대한 검정은 하우스만 검정(Hausman Specification Test)이다.

제3장 한국복지패널을 이용한 패널자료분석

3.1 한국복지패널 조사

한국복지패널(KWPS : Korea Welfare Panel Study)은 보건복지부, 한국보건사회연구원, 서울대학교의 논의에 의해 기존에 조사되고 있던 한국보건사회연구원의 차상위·빈곤패널과 자활패널, 서울대의 복지패널을 통합하여 2006년도부터 한국보건사회연구원의 주관으로 조사가 진행되고 있다.

한국복지패널의 목적은 외환위기 이후 빈곤층, 근로빈곤층(working poor), 차상위층(near poor)의 가구형태, 소득수준, 취업상태가 급격히 변화하고 있어, 이들의 규모와 실태변화를 동태적으로 파악하여 정책지원을 위한 기초 자료를 생산하고, 소득계층별, 경제활동상태별, 연령별 등 각 인구집단의 생활실태와 복지욕구 등을 역동적으로 파악하고 정책효과성을 평가함으로써 정책형성과 피드백에 기여하고자 하는 것이다. 따라서 조사대상 가구를 일반가구와 저소득층 가구로 분류하고 각각 50%씩 추출하여 이들에 관한 다양한 복지실태와 욕구 등을 조사하였다.

<표 3-1> 한국복지패널 조사의 개요

구 분	인구주택 총조사(90%)	2006 국민생활실태조사	한국복지패널 조사
조 사 구	23만여개 조사구	517개 조사구	446개 조사구
가 구 수	14백만여 가구	30,000가구	7,000가구
추출방법		2단계층화집락	층화집락계통
대 표 성		전국	전국

<표 3-1>을 살펴보면, 2005년도 인구주택 총조사 90%자료로부터 2006 국민생활실태조사가구 30,000가구를 2단계 층화집락추출에 의해 추출하였고, 이들 가구 중 소득계층별로 저소득층 가구와 일반가구를 각각 3,500가구씩을 층화집락계통추출을 통해 총 7,000가구를 표본으로 선정하여 최종 7,072가구를 조사하였다(한국보건사회연구원, 2007).

연도별로 조사 현황을 살펴보면, 2006년 1차년도 패널에는 총 7,072가구, 14,469명을 조사 완료하였으며, 2007년에는 총 6,580가구, 13,480명을 그리고, 2008년에는 총 6,314가

구, 12,930명을 조사 완료하였다. 1차년도 대비 2차년도, 3차년도 조사완료 가구 및 가구원 비율을 살펴보면 가구는 각각 93.04%, 89.28%, 가구원은 각각 93.15%, 89.36%로서 전체적으로 1차년도 대비 89%이상의 완료율을 보였다.

<표 3-2> 한국복지패널의 연도별 자료 현황

구 분	2006년			2007년			2008년		
	일반 가구	저소득 가구	합 계	일반 가구	저소득 가구	합 계	일반 가구	저소득 가구	합 계
가구수	3,789	3,283	7,072	3,755	2,825	6,580	3,695	2,619	6,314
가구원수	8,646	5,823	14,469	8,651	4,829	13,480	8,544	4,386	12,930

3.2 분석목적 및 이론적 배경

한국복지패널의 조사대상 중 저소득가구(low-income)는 표본의 50%를 할당할 정도로 중요한 관심사이며, 1960년대 이후 급격한 경제성장에서 발생하는 사회적 관심사 중의 하나이다. 이에 본 논문에서는 저소득가구의 삶의 질에 대해 연구하고자 하였으며, 이를 위해 저소득가구의 삶의 질에 영향을 미치는 요인들이 무엇인지에 대해 살펴보았다. 또한 일반가구의 삶의 질에 영향을 미치는 요인들과 비교함으로써 저소득가구와 일반가구 간의 차이를 제시하고자 하였다.

일반적으로 저소득층(low-income)은 경제적 빈곤과 관련하여 최저생활을 유지하는데 필요한 소득이 안 되는 상태의 객관적인 빈곤선을 설정하여 여기에 못 미치는 개인이나 가구를 의미한다. 1960년대 이후 급격한 경제성장의 결과 전반적 생활수준의 향상 및 물질적 풍요로 인해 절대빈곤이 어느 정도 해소되었지만, 경제발전과정 중에서 부의 불공정한 분배 및 빈부격차 등으로 상대적 빈곤문제가 더욱 심화되고 있으며, 절대빈곤으로 어려움을 겪고 있는 층이 아직도 존재하고 있다(김혜묵, 1999).

저소득층은 빈곤계층과 유사하게 사용되기도 하며 개념에 대해 학자들마다 다양하게 분류하고 있다. 사회적 통념으로 볼 때 저소득층은 기초생활 보호대상자뿐만 아니라 대개 빈곤자를 포함하는 경향이 있다. 즉 저소득계층은 영세민을 포함하여 불안정 취업으로 생활하는 계층을 총괄하는 개념이다.

저소득층에 대한 개념을 종합적으로 정리해 보면, 저소득층은 경제적, 사회적으로 매우

유동적인 상황에 처해 있는 주변적인 존재(marginal man)이며, 소득, 주거환경, 취업구조, 사회적 자원의 배분양식 등 다양한 요인들이 복합적으로 얽히면서 그 독특한 성격이 규정됨을 알 수 있다(박성, 2003). 또한 저소득층의 규정을 기초생활보호대상자에 속하는 자활보호대상자로서 대개 활동능력은 있지만 실직상태에 놓여있거나, 취업을 하더라도 취업구조에서 낮은 임금과 취업의 불안정성으로 자립을 못하고 있는 자로 규정하면서, 공식적, 비공식적 부문이나 영세자영업, 소규모 서비스업계 등에 일시적으로 고용된 노동력으로 정의하였다. 특히 저소득층은 열악한 주거환경 속에서 저소득자의 불안전 취업이라는 특징을 공유하면서 도시외곽이나 도심의 빈곤지역에 밀집되어 살아가고 있는 가난한 사람들로 규정하였다(박용순, 2001).

다음으로 삶의 질의 개념에 대해 살펴보면, 삶(life)이란 살아가는 일, 살아가는 현상, 목숨 또는 생명을 지칭하며, 질(quality)이란 어떠한 것이 소유한 우수성의 정도나 어떤 대상을 좋거나 나쁘게 할 수 있는 것이다.

Dubos(1976)는 일상생활의 활동에서 얻는 만족감과 관계되는 주관적인 가치판단이라 하였고, George(1980)은 삶의 만족, 자아존중감, 건강상태와 기능 및 사회·경제 상태에 대한 주관적·객관적 평가로 삶의 질을 정의하였다. 또한 이들 삶의 전반적인 상황이나 삶의 경험들에 대한 개인의 주관적인 평가와 만족으로 정의하기도 하였다(Magilvy, 1985).

한국보건사회연구원에서의 '한국인의 삶의 질과 과제'(1997)에서는 삶의 질이란 인간생활의 질적 수준을 나타내는 의미로 개인의 주관적 만족도 또는 행복감을 뜻하는 포괄적인 의미를 갖는 것으로 삶의 질과 유사한 개념으로 안녕(well-being), 복지(selfare), 생활만족도(life satisfaction), 행복감(happiness), 사회지표(social indicator), 생활수준(standard of life), 주관적 안녕감(subjective well being), 생활의 질로 표현하였다.

즉, 삶의 질이란 신체·정신적 및 사회·경제적 영역에서 각 개인이 지각하는 주관적 안녕감을 의미하는 것으로(노유자, 1988), 관련 선행연구들의 결과를 통해 삶의 질에 영향을 미치는 요인에 대해 살펴보면 다음과 같다.

김은주(2001)의 연구에서는 삶의 만족도가 연령, 교육수준, 종교, 배우자 유무, 자녀유무에서 차이를 보이지 않았으나 일상생활 수행정도는 부분적으로 차이가 있는 것으로 나타났다. ADL과 IADL에 따른 삶의 만족도에서 활동 장애가 없을수록 삶의 만족도가 높았다. 일생생활 능력정도는 개인의 독립적이며 의미 있는 생활을 유지하고 나아가 자기관리의 자립에 영향을 미치므로 활동장애가 없는 사람일수록 삶의 만족도는 당연히 높아질 것이다.

서영희(1994)는 자아존중감이 삶의 질에 영향을 주는 요인으로 설명하였다. 박은숙(1998)은 연령, 경제상태, 교육정도 등의 인구사회학적 변수와 건강증진 행위 및 건강개념,

자아존중감, 내적 건강통제, 자기효능감 등의 변수가 노인의 삶의 질을 예측하는 변수라고 하였으며, 그 중 노인의 삶의 질에 가장 큰 영향을 주는 변수는 자아존중감이라고 하였다.

3.3 분석방법 및 변수

본 논문에서는 한국복지패널을 활용하여 저소득가구의 생활만족도에 영향을 미치는 요인들이 무엇인가 그리고 일반가구와의 차이점은 무엇인가에 대해 살펴보았다.

이상의 통계분석은 SAS 9.1.3 프로그램의 PROC MIXED과 PROC TTEST를 사용하였다. PROC MIXED는 혼합모형의 분석에 사용되는 도구이지만 PROC TSCSREG와 함께 패널모형 분석에서도 사용되고 있다. 그 이유를 살펴보면, 패널모형에서는 상수항에 대해서만 확률효과를 가정하고 있다. 반면 혼합모형에서는 상수항뿐만 아니라 회귀계수에도 확률효과를 가정할 수 있다. Free(2004)는 패널모형과 같이 상수항에 확률효과를 가정한 모형을 오차성분모형(error component model) 또는 확률절편모형(random intercepts model)이라 하고, 회귀계수에 확률효과를 가정한 모형을 확률계수모형(random coefficients model)이라 정의하고 있다. 확률효과가 상수항에 제한되어 있는 패널모형이 혼합모형의 하위개념이라고 볼 수 있다.

<표 3-3> 변수의 정의

구 분		변 수	척 도	내 용
종 속 변 수		전반적만족도	순위형	매우 불만족. 대체로 불만족, 그저 그렇다, 대체로 만족, 매우 만족
독립변수	인구학적 특성	성 별	명목형	여성, 남성
		연 령	연속형	
		거주지역	명목형	서울시, 광역시, 기타지역
		교육수준	순위형	초졸이하, 중졸, 고졸, 대졸이상
		종교유무	명목형	종교 유, 종교 무
		직 업	명목형	전문직, 사무직 및 판매·서비스업, 농·어·수산업 및 기능직, 단순노무직
	가족 자원	가구원수	연속형	
		결혼여부	명목형	기혼, 미혼
		가족원간 의견 충돌	순위형	전혀 그렇지 않다, 그렇지 않은 편이다, 보통이다, 그런 편이다, 매우 그렇다
	경제 자원	소 득	연속형	가처분소득
		주거의 점유형태	명목형	자가, 전세, 보증부월세, 월세, 기타
	건강 자원	흡연여부	명목형	흡연 유, 흡연 무
		1년 평균 음주량	순위형	전혀 마시지 않는다, 주 1회 이하, 주 2~3회, 주 4회 이상
		건강상태	순위형	건강이 아주 안 좋다, 건강하지 않은 편이다, 보통이다, 건강한 편이다, 아주 건강하다
	자아 존중감	긍정적 자아존중감	연속형	Factor1
		부정적 자아존중감	연속형	Factor2

관련 선행연구에서의 삶의 질에 대한 개념은 삶의 상황이나 경험에 대한 주관적 평가와 만족 혹은 행복, 신체적, 정서적, 사회적으로 안녕한 상태, 삶의 조건에 대한 주관적인 만족 상태 등에 따라 다양하게 존재하지만 본 논문에서는 생활실태·만족 및 의식을 묻는 문항에서 전체적인 만족도의 정도를 종속변수로 정의하였다. 조사 문항은 '건강, 가족의 수입, 주거 환경, 가족 관계, 직업, 사회적 친분관계, 여가생활의 사항을 모두 고려할 때, 귀하는 전반적으로 생활에 얼마나 만족하고 계십니까?'에 대한 5점 척도로 '매우 불만족', '대체로 불만족', '그저 그렇다', '대체로 만족', '매우 만족'으로 하여 점수가 높을수록 생활만족도가 높음을 의미한다.

독립변수는 선행연구를 바탕으로 하여 성별, 연령, 거주 지역, 교육수준, 직업, 종교 유무 등의 "인구학적 특성변수", 가구원 수, 결혼여부, 가족 간의 의견 충돌 정도 등의 "가족자원 변수", 가처분소득, 주거의 점유형태 등의 "경제적 자원 변수", 흡연여부, 1년 평균 음주량, 건강상태 등의 "건강자원 변수", 마지막으로 긍정적 자아존중감과 부정적 자아존중감 등의 "자아존중감 변수"를 선택하여 분석하였다. 독립변수에 대한 구체적 내용은 <표 3-3>과 같다.

일반적으로 독립변수가 많아질수록 다중공선성의 문제가 심각해지는 경향을 갖는다. 변수의 수가 많아지면 자유도의 손실이 커지므로, 의미 있는 추정을 수행하기 어려워지고 계수값의 신뢰도가 떨어진다. 또한 변수 선택의 간편성(parsimoniousness)측면에서 볼 때도 변수가 많은 것이 좋은 모형이 아님을 고려해 볼 때 적절히 변수를 축소시킬 필요가 있다(이종원, 2001). 이러한 상황에서 본 논문에서는 자아존중감과 관련된 변수들에 대하여 변수들이 가지고 있는 특징을 보존시키면서 소수의 새로운 변수를 생성하여 패널 데이터 분석을 수행하고자 하였다.

자아존중감은 <표 3-4>의 총 10개의 문항에 대해 응답자들이 지각하는 정도를 '매우 불만족', '대체로 불만족', '그저 그렇다', '대체로 만족', '매우 만족'의 5점 척도로 측정하였다. 본 논문에서는 전체 10개의 문항에 대하여 요인분석(factor analysis)을 실시하고, 요인점수(factor score)를 독립변수로 활용하였으며 그 결과는 <표 3-4>와 같다. 또한 주성분분석법(principal components method)을 이용하여 2개의 요인을 추출하였으며 베리맥스 회전(varimax rotation)을 이용하였다. 추출된 2개 요인은 각각 긍정적 자아존중감(factor1), 부정적 자아존중감(factor2)이라 명명하여 독립변수로 사용하였다.

<표 3-4> 자아존중감에 대한 요인분석

자 아 존 중 감	Factor1	Factor2
나는 대부분의 다른 사람들처럼 일을 잘할 수 있다	0.70583	-0.21516
나는 내 자신에 대해 긍정적인 태도를 가지고 있다	0.68989	-0.29123
나는 내가 가치 있는 사람이라고 생각한다	0.66459	-0.35640
나는 좋은 성품을 가졌다고 생각한다	0.60823	-0.15129
나는 내 자신에 대하여 대체로 만족한다	0.60590	-0.47137
나는 내 자신을 좀 더 존경했으면 좋겠다	0.52442	0.16099
나는 대체적으로 실패한 사람이라고 생각한다	-0.11335	0.76542
나는 가끔 내 자신이 쓸모없는 사람이라는 느낌이 든다	-0.17782	0.70813
나는 자랑할 것이 별로 없다	-0.10215	0.69337
나는 때때로 내가 좋지 않은 사람이라고 생각한다	-0.13008	0.43896

다음으로 각 요인 안에 포함되어 있는 변수들이 내적 일관성이 있는지에 대해 알아보았다. 한 개념을 많은 항목으로 측정했을 때 그 항목들에 대한 일관성을 측정하는 것으로 크론바하 알파(Cronbach's α)계수를 이용한다. 이 값은 0에서 1사이의 값을 가지며 높을수록 좋으나, 0.6이상의 값을 가지면 수용할 정도의 수준이라고 알려져 있다. 긍정적 자아존중감(factor1)에 포함된 문항들에 대한 크론바하 알파 계수는 0.7623이며, 부정적 자아존중감(factor2)에 포함된 문항들에 대한 크론바하 알파 계수는 0.6547이다. 따라서 이 두 요인 모두 내적 일관성이 있다고 할 수 있다.

3.4 분석결과

본 논문에서는 생활만족도와 그에 영향을 미치는 요인들이 저소득가구와 일반가구 간에 차이가 나타날 것으로 생각하였다. 이러한 차이점을 확인하기 위해 먼저, 한국복지패널에서 건강 만족도, 가족의 수입 만족도, 주거 환경 만족도, 가족관계 만족도, 직업 만족도, 사회적 친분관계 만족도, 여가 생활 만족도, 전반적 만족도 등의 생활만족도에 대한 문항들에 대해 t-검정을 실시하였다.

전반적 만족도를 제외한 나머지 7개 만족도의 t-검정결과를 보면, 저소득가구와 일반가구 간에는 각각의 만족도에서 차이가 있는 것으로 나타났다.

<표 3-5> 저소득가구와 일반가구 간의 각 문항별 만족도에 대한 t-검정결과

구분	저소득가구		일반가구		t값
	평균	표준편차	평균	표준편차	
건강 만족도	2.537	1.138	3.349	1.002	70.57***
가족의 수입 만족도	2.154	0.862	2.835	0.922	73.03***
주거 환경 만족도	3.197	0.977	3.450	0.861	25.43***
가족 관계 만족도	3.559	0.866	3.977	0.699	49.01***
직업 만족도	2.693	0.961	3.239	0.946	54.28***
사회적 친분관계 만족도	3.499	0.850	3.781	0.711	33.38***
여가 생활 만족도	2.668	0.942	3.021	0.960	34.28***

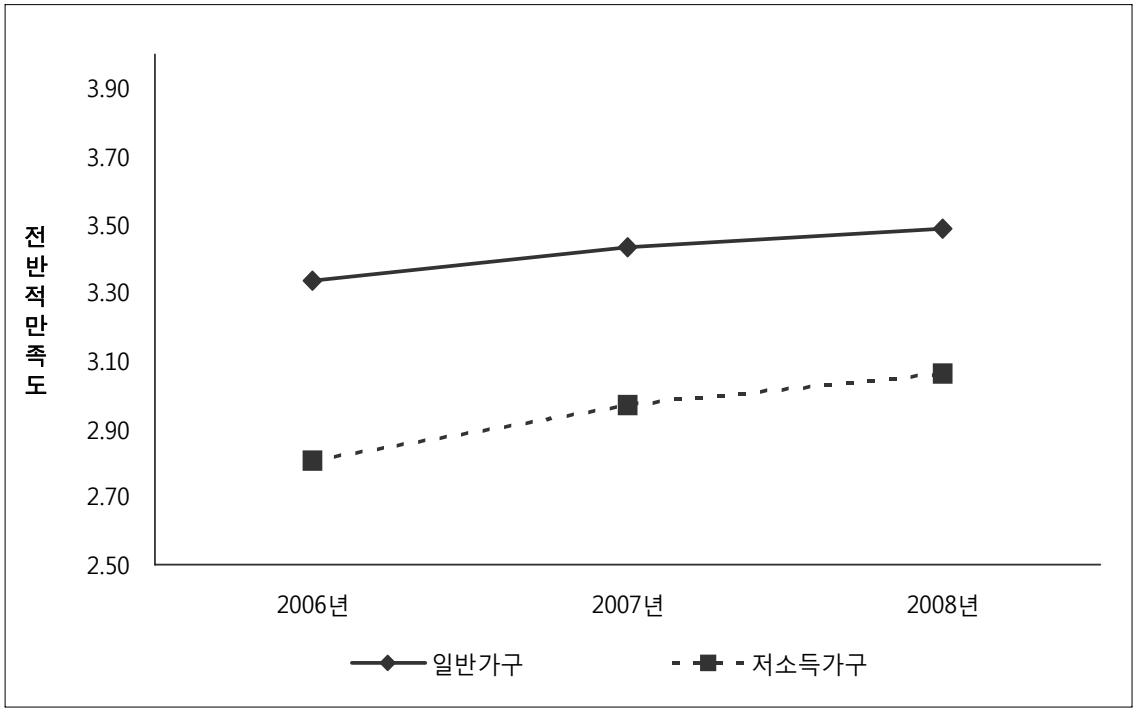
*** p<0.01, ** p<0.05, * p<0.1

전반적 만족도는 각 연도별 및 전체에 대해 살펴보았다. 그 결과, 각 연도별 및 전체에서 모두 저소득가구와 일반가구 간에 전반적 만족도에서 유의한 차이가 있는 것으로 나타났다. 또한, <그림 3-1>에서도 각 연도별로 저소득가구와 일반가구 간에 전반적 만족도의 평균이 차이가 있음을 확연히 보여주고 있다.

<표 3-6> 저소득가구와 일반가구 간의 전반적 만족도에 대한 t-검정결과

구분	저소득가구		일반가구		t값
	평균	표준편차	평균	표준편차	
2006년	2.808	0.780	3.333	0.670	40.33***
2007년	2.967	0.803	3.435	0.727	32.50***
2008년	3.064	0.766	3.487	0.693	29.86***
전　체	2.933	0.791	3.418	0.710	60.25***

*** p<0.01, ** p<0.05, * p<0.1



<그림 3-1> 저소득가구와 일반가구의 연도별 전반적 만족도 평균

이상의 결과를 통해 저소득가구와 일반가구 간에는 생활만족도에서 차이가 있음을 알 수 있으며, 본인의 삶에 만족하는 정도는 저소득가구에 비해 일반가구가 더 높다는 것을 알 수 있다. 이러한 결과를 바탕으로 본 논문에서는 저소득가구와 일반가구 간에 생활만족도뿐만 아니라 그에 영향을 미치는 요인에도 차이가 있을 것으로 생각하였다. 하지만, t-검정의 결과는 저소득가구와 일반가구 간에 생활만족도 차이는 비교할 수 있으나, 어떠한 요인이 이들의 생활만족도에 영향을 미치면서 어느 정도의 영향력을 가지고 있는지는 알 수 없다. 따라서 추가적인 분석방법이 필요하다.

이를 살펴보기 위해 한국복지패널에서 저소득가구와 일반가구에 대한 각각의 패널모형을 설정하였으며, 그 차이를 살펴보기 위해 독립변수는 동일하게 선택하였다. 이 때 패널모형은 각각 시간효과를 랜덤으로 하는 1요인 랜덤효과 모형을 이용하였다.

저소득가구와 일반가구의 생활만족도에 영향을 미치는 요인에 대한 분석결과는 <표 3-7>에서 제시하였다. 그 결과를 살펴보면, 인구학적 특성에서는 '성별'과 '직업', 가족자원에서는 '가구원수'와 '가족원간 의견 충돌 정도', 경제지원에서는 '주거상태', 건강자원에서는 '건강상태', 자아존중감은 '긍정적 자아존중감'과 '부정적 자아존중감'이 두 가구에서 공통적

으로 유의하게 생활만족도에 영향을 미치는 요인으로 나타났다.

<표 3-7> 생활만족도에 영향을 미치는 요인에 대한 분석결과

구 분		저소득가구				일반가구			
		추정치	표준오차	t값	F값	추정치	표준오차	t값	F값
절 편		2.654	0.156	17.1***		2.949	0.077	38.6***	
인구학적 특성	성별	-0.059	0.031	-1.9*	3.5*	-0.047	0.016	-2.9***	8.8***
	연령	0.001	0.001	0.5	0.3	-0.001	0.001	-1.5	2.3
	지역(1)	-0.043	0.035	-1.2	0.7	-0.041	0.015	-2.7***	5.0***
	지역(2)	-0.014	0.029	-0.5		-0.036	0.014	-2.5**	
	교육수준	0.009	0.018	0.5	0.2	0.042	0.009	4.6***	21.4***
	종교유무	-0.032	0.026	-1.3	1.6	-0.022	0.012	-1.8*	3.1*
	직업(1)	0.175	0.064	2.8***	3.3**	0.210	0.022	9.5***	33.8***
	직업(2)	-0.022	0.030	-0.7		0.047	0.018	2.7***	
	직업(3)	0.013	0.040	0.3		0.069	0.018	3.8***	
가족자원	가구원수	-0.025	0.013	-1.9**	3.9**	-0.032	0.006	-5.5***	30.4***
	결혼여부	-0.042	0.030	-1.4	1.9	-0.052	0.016	-3.2***	10.3***
	가족충돌	-0.048	0.011	-4.6***	20.9***	-0.052	0.006	-9.1***	83.5***
경제자원	소득	-0.001	0.001	-0.1	0.0	0.001	0.001	7.7***	59.2***
	주거(1)	0.134	0.045	2.9***	15.3***	0.031	0.028	1.1	27.9***
	주거(2)	-0.004	0.051	-0.1		-0.075	0.030	-2.5**	
	주거(3)	-0.072	0.049	-1.5		-0.126	0.030	-4.2***	
	주거(4)	-0.199	0.065	-3.1***		-0.187	0.049	-3.9***	
건강자원	흡연여부	0.036	0.033	1.1	1.2	0.102	0.015	6.7***	44.9***
	음주량	-0.006	0.010	-0.6	0.3	0.001	0.005	0.3	0.1
	건강상태	0.144	0.014	10.4***	107.9***	0.101	0.008	13.2***	174.9***
자아존중감	긍정적	0.138	0.013	10.6***	112.0***	0.146	0.007	22.3***	497.5***
	부정적	-0.240	0.012	-19.6***	384.3***	-0.201	0.007	-27.7***	765.8***

*** p<0.01, ** p<0.05, * p<0.1

반대로 저소득가구와 일반가구를 비교해 보면, 일반가구에서는 거주 지역, 교육수준, 종교 유무, 결혼 여부, 소득, 흡연 여부 등이 생활만족도에 영향을 주는 반면, 저소득가구에는 이러한 요인들이 영향을 주지 않는 것으로 나타났다. 이 결과에서는 생활만족도에 영향을 주는 요인의 수가 저소득가구의 경우가 일반가구에 보다 다소 적다는 것이 큰 특징이다. 즉, 대체로 생활만족도는 저소득가구가 일반가구에 비해 낮지만, 일반가구의 경우 생활만족도에 영향을 미치는 요인이 저소득가구에 비해서는 다양하다는 것이다.

저소득가구의 결과를 좀 더 살펴보면, 먼저 인구학적 특성에서는 남성보다는 여성의 경우, 다른 직업보다 전문직일 경우에 생활만족도가 높은 것으로 나타났다. 가족자원에서는 가구원의 수가 작고 가족원간 의견 충돌 정도가 작을수록 생활만족도가 높으며, 경제자원에서는 본인 소유의 집이 있는 경우에 생활만족도가 높은 것으로 나타났다. 건강자원과 자아존중감에서는 건강강태가 좋으면서 긍정적 자아존중감이 높고 부정적 자아존중감은 낮을수록 생활만족도가 높은 것으로 나타났다.

이러한 결과들은 선행연구의 결과들과도 일치하는 것으로 저소득가구의 경우 일반가구에 비해 생활만족도는 낮은 편이며, 저소득가구 내에서도 전문직이거나 본인 소유의 집이 있고 긍정적 자아존중감이 높고 부정적 자아존중감은 낮을수록 생활만족도가 높은 것을 알 수 있다.

제4장 결론 및 향후과제

4.1 결 론

본 연구에서는 횡단면적인 정보뿐만 아니라 시계열 정보까지 보유하고 있는 패널자료에 대해 살펴보고 이러한 자료를 분석할 수 있는 패널모형과 분석방법에 대해서도 살펴보았다. 또한, 실증분석을 위해 한국보건사회연구원의 주관으로 조사된 한국복지패널(KWPS : Korea Welfare Panel Study)자료를 SAS 9.1.3의 PROC MIXED을 사용하여 분석하였다.

한국복지패널은 외환위기 이후 빈곤층, 근로빈곤층(working poor), 차상위층(near poor)의 가구형태, 소득수준, 취업상태가 급격히 변화하고 있어, 이들의 규모와 실태변화를 동태적으로 파악하여 정책지원을 위한 기초 자료를 생산하고, 소득계층별, 경제활동상태별, 연령별 등 각 인구집단의 생활실태와 복지욕구 등을 역동적으로 파악하고 정책효과성을 평가함으로써 정책형성과 피드백에 기여하고자 하는 목적으로 조사된 자료이다.

한국복지패널 자료 중 생활실태·만족 및 의식을 묻는 문항에서 전반적 만족도를 종속변수로 정의하고 독립변수는 성별, 연령, 거주 지역, 교육수준, 직업, 종교 유무의 “인구학적 특

성변수", 가구원 수, 결혼여부, 가족 간의 의견 충돌 정도의 "가족자원 변수", 가처분"가족 자주거의 점유형태의 "경제적 자원 변수", 원 여부, 1년 평균 음주량, 건강상태의 "건강자원 변수", 긍정적 및 부정적 자아존중감의 "자아존중감 변수"를 선택하였다. 또한, 종속 및 독립변수를 분석하기 위해 1요인 랜덤 효과 모형을 설정하였다.

모형구축에 앞서 자아존중감의 총 10개 문항에 대해 요인분석(Factor analysis)을 실시하였으며, 요인점수(factor score)를 이용하여 10개의 문항을 각각 긍정적 자아존중감과 부정적 자아존중감으로 나누었다.

또한 t-검정을 통해 저소득가구와 일반가구 간의 생활만족도의 차이를 살펴보았으며 그 결과, 일반가구가 저소득가구보다 생활만족도가 더 높은 것으로 나타났으며 그 차이는 통계적으로 유의하다는 것을 확인하였다.

저소득가구와 일반가구의 생활만족도에 영향을 미치는 요인에 대해 분석한 결과에서는 일반가구의 경우에는 연령과 1년 평균 음주량만이 생활만족도에 영향을 미치지 않는 반면, 저소득가구에서는 연령과 1년 평균 음주량 이외에도 거주 지역, 교육수준, 종교 유무, 결혼 여부, 소득, 흡연 여부 등의 변수가 생활만족도에 영향을 미치지 않았다.

이러한 결과를 종합해 보면, 본 논문의 목적 중 하나인 저소득가구의 삶의 질에 대한 연구결과는 먼저 인구학적 특성에서는 남성보다는 여성의 경우, 다른 직업보다 전문직일 경우에 생활만족도가 높고 가족자원에서는 가구원의 수가 작고 가족원간 의견 충돌 정도가 작을수록 생활만족도가 높으며, 경제자원에서는 본인 소유의 집이 있는 경우에 생활만족도가 높은 것으로 나타났다. 건강자원과 자아존중감에서는 건강강태가 좋으면서 긍정적 자아존중감이 높고 부정적 자아존중감은 낮을수록 생활만족도가 높은 것으로 나타났다.

4.2 향후과제

본 연구에서 실시한 패널분석은 저소득가구와 일반가구 간의 생활만족도에 영향을 미치는 요인은 비교할 수 있으나 어떠한 요인이 이들의 생활만족도에 어느 정도의 영향력을 가지는지 알 수 없다. 따라서 보다 구체적인 영향정도를 알기 위해서는 보다 다양한 분석이 요구된다. 또한, 패널모형 중 1요인 랜덤 효과 모형을 사용하였으나 추가적으로 다른 모형들과의 비교를 통해 가장 적합한 모형을 찾는 과정이 필요하다.

추가적으로 패널자료는 각 관측대상의 관측횟수에 따라 균형패널과 불균형패널로 구분되는데, 조사대상자들의 무응답 또는 조사거부 등으로 인해서 균형패널을 구성하기가 어려운 것이 현실이다. 또한 모형적합을 위한 데이터셋 생성과정에서 해당 연도에 조사가 되었지

만, 모형에서 사용된 일부 문항들이 조사가 되지 않아서 결측값으로 처리되어 제거되는 관측값들도 적지 않다. 따라서 횡단면 대체(cross-sectional imputation)와 경시적 대체(longitudinal imputation)로 구분된 패널자료의 결측값 대체 방법을 적용하여 정보의 손실을 최소화하는 방안이 필요하다.

참고문헌

[1] 김은주. (2001). 시설노인과 재가노인의 일상생활정도와 생활만족도의 비교, 청주대학교 석사학위논문

[2] 김혜묵. (1999). 저소득 노인의 여가실태에 관한 연구 - 강서구 영구임대아파트를 중심으로, 단국대학교 석사학위논문

[3] 노유자. (1988). 서울지역 중년기 성인의 삶의 질에 관한 분석 연구, 연세대학교 박사학위논문

[4] 박성. (2003). 도시저소득층 노인의 여가생활과 개선방안에 관한 연구, 서울시립대학교 석사학위논문

[5] 박용순. (2001). 노인자원봉사의 활성화를 위한 실증적 분석연구, 한국사회복지학회, 46

[6] 박은숙, 김순자, 김소인, 전영자, 이평숙, 김행자, 한금선. (1998). 노인의 건강증진 행위 및 삶의 질에 영향을 미치는 요인, 고려대 간호학논집, 1

[7] 서영희. (1994). 사회적 기대 지각과 스트레스간의 관계, 효성여자대학교 석사학위논문

[8] 안두진. (2008). SAS PROC PANEL을 이용한 패널자료분석, 숭실대학교 석사학위논문

[9] 이영훈. (2001). 선형패널자료 모형에 관한 문헌연구, 계량경제학보, 15(1), pp.105-138, 한국계량경제학회

[10] 이종원. (2001). 계량경제학, 박영사, pp.334-335

[11] 윤기윤. (2007). 저소득 독거노인의 삶의 질에 영향을 미치는 요인에 관한 연구, 국제신학대학원대학교 박사학위논문

[12] 전승훈, 강성호, 임병인. (2004). 선형패널자료 분석방법에 관한 비교연구, 통계연구, 제9권 제2호, pp.1-24, 통계청

[13] 최충익. (2004). 도시화에 따른 수해 취약성에 관한 실증분석: 경기도 패널데이터를 활용하여, 국토연구, 제42권, pp.17-37

[14] 최충익. 패널모형: 시계열 분석과 횡단면 분석을 한번에, 알기 쉬운 연구방법론(28)

[15] 한국보건사회연구원. (1997). 한국인의 삶의 질과 과제

[16] 한국보건사회연구원. (2007). 2007 한국복지패널 기초분석 보고서

[17] Ashenfelter, O., Levine B. P. and Zimmerman J. D. (2003), Statistics and Econometrics: Methods and Applications, John Wiley & Sons. Inc.

[18] Baltagi, B. H. (2005). Econometric Analysis of Panel Data, Third Edition,

Wiley, New York

[19] Free, E. W. (2004). Longitudinal and Panel Data : Analysis and Applications in the Social Sciences, Cambridge : Cambridge University Press

[20] Hsiao, C. (2003). Analysis of Panel Data, Second Edition, Cambridge University Press

[21] Jonhston, J. and DiNardo, J. (1997). Econometric Methods, Mcgraw-Hill International Editions

[22] Magilvy. (1985). Quality of Life Hearing-Impaired Older Women, Nursing Research, 34(2), pp.140-144

[23] R. Dubos. (1976). The State of Health and The Quality of life, The Western Journal of Medicine, 125(1), pp.8-9

[투고일: 2009. 08. 10][심사(수정)일: 2009. 08. 15][게제확정일: 2009. 08. 19]

Annals of Statistical Research
Vol. 6. No. 1, 2009, pp.27~40

The study of causes of cancelling long-term accident compensation insurance

Sun Jung[1], Min-Su Kang[2], Hye-Rim Lee[3]

Abstract

After 1950's, most of insurance companies in Korea have been developed their system in aspects of both qualities and quantities. In this condition, many people take insurance service for not only preventing their loss but also providing for unforeseen accidents. In these days, carrying insurances become the one of plenty of ways to reserve one's properties and financial technology. However, the risk according to insurance cancellations is also rising. Then, we need to study the causes of insurance cancellations, especially, long-term accident compensation insurance cancellation. For this study, we use statistical methods, logistic regression model and ROC curves with cross validation.

Key words : long-term accident compensation insurance, logistic regression model, cross validation, ROC curve

1) Doctoral Candidate for Statistics, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : iamlucky@dongguk.edu
2) Graduate student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : minsu7733@hanmail.net
3) Graduate student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : leehr26@dongguk.edu

Ⅰ. 서 론

자본주의 사회에서 같은 종류의 사고를 당할 위험성이 있는 많은 사람이 미리 금전을 각출하여 공통준비재산을 형성하고, 사고를 당한 사람이 이것으로부터 재산적 급여를 받는 경제제도를 보험이라 한다. 한국 보험의 뿌리는 삼한시대 계(契)였으며, 근대적 보험은 일제 강점기인 1921년, 1922년에 설립된 조선생명보험주식회사, 조선화재해상보험주식회사가 있었으나 보험 대상자는 사람이 아닌 소였다. 광복 이후 일본 생명 보험사들이 환급금 보상 없이 철수하였고, 한국전쟁 후 1950년대 대한생명이 우리나라 최초의 보험사로 창업 된 이래 1960년 경제개발 추진으로 생명 보험사가 국민 저축 기관으로 지정되어 1970년 경제성장에 힘입어 발전하였다 (이경룡, 2002). 보험은 새마을 운동의 일환으로 초기에는 저축성 상품과 교육보험위주의 상품으로 국민들 가정에 파고 들어가기 시작한 보험은 지금은 없어서는 안 될 생활 필수품이 되었다.

우리나라 보험의 흐름은 크게 3세대로 살펴볼 수 있다. 1977년 보험의 해를 기준으로 1998년 IMF 금융위기까지를 1세대로 이 시기에는 보험은 보장을 중요시하였다. 노인문제가 사회 문제로 대두하는 1999년부터 2003년은 보험의 2세대로 IMF이후 발생한 대량실직, 조기 은퇴, 고용불안, 효도를 기대할 수 없는 전통 문화 붕괴, 가족 문화 쇠퇴 등으로 위기의식이 팽창하면서 종신 보험이 기본 상품이 되어가고, 고령화로 인한 노후 준비를 서두르게 되었다 (허연, 2006). 제 2세대의 보험은 보장과 더불어 노후준비의 수단으로 변해가고 있다. 3세대의 보험은 2004년부터 시작된 세계적인 저금리 현상과 물가 상승으로 인해 변화가 시작된다. 이 시에는 단순한 보장만을 위한 보험의 아니라 투자 개념이 가미된 변액 보험을 통해 이제는 보험은 재테크의 수단으로 자리 잡아가고 있다.

보험의 다양한 보장과 발전에도 불구하고 다수의 보험계약자들은 많은 액수의 보험료를 지불하고서도 제대로 보장받고 있는지 조차 모르고서 일을 당한 뒤에야 보험회사를 원망하는 경우가 많이 발생한다. 아는 만큼 보인다는 말처럼 보험의 본 연구에서는 다양한 보험 피해 중에서 중도 해지에 따른 위험을 줄여보고자 보험 상품의 해지에 영향을 미치는 요인들을 알아보려한다.

본 논문의 Ⅱ장에서는 보험 해지에 영향을 미치는 요인을 찾기 위해 로지스틱 회귀분석모형과 설정된 모형을 평가하는 방법으로 교차타당성 방법과 ROC곡선 및 장기손해보험에 대한 이론적 근거를 제시한다. 그리고 A보험회사의 장기손해보험상품 자료를 이용하여 로지스틱 회귀분석을 통한 모형을 설정하고, 이를 평가하여 이 상품의 해지에 영향을 주는 요인들을 보이고 설정된 모형을 평가한다. 마지막 Ⅲ장에서는 본 논문의 결론과 아직 수행하지

못한 향후 연구 과제를 제시한다.

Ⅱ. 본론

본 연구에서는 A보험회사의 장기손해보험상품의 자료를 로지스틱 회귀분석(Logistic regression) 방법을 이용하여 모형을 설정하고, 설정된 모형을 평가하기 위해서 교차타당성(Cross-validation) 방법과 ROC 곡선(The Receiver Operating Characteristic Curve)을 사용한다. 다음은 본 연구에서 사용된 로지스틱 회귀모형과, 교차 타당성 방법과 ROC 곡선 및 장기손해보험의 이론적 근거이다.

1. 로지스틱 회귀모형 (Logistic regression model)

선형 회귀모형에서는 독립변수와 종속변수 모두를 수치형 변수로 가정한다. 그러나 일반적으로 데이터에는 성별과, 직업 등 수치형 변수가 아닌 범주형 변수가 존재하기 마련이다. 특히, 종속변수가 범주형 변수로 표현되는 회귀모형을 로지스틱 모형이라고 한다. 선형 회귀모형에서는 오차항의 정규성, 등분산성, 독립성의 가정이 필요하지만, 로지스틱 모형에서는 반응함수가 확률로써 정의되기 때문에 이로 인한 범위의 제약으로 선형 회귀모형에서의 가정은 적용할 수 없게 된다 (염준근, 2005). 범주형 자료에서 가장 일반적인 형태로써 종속변수가 이항자료(binary data)일 경우(이항반응변수, Y)의 로지스틱 함수는 다음과 같은 식으로 표현할 수 있다.

$$y = \log it(\pi(x)) = \log(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x \qquad (1)$$

이항반응변수 Y($Y=1$일 때는 성공, $Y=0$일 때는 실패)와 설명변수 X를 통한 로지스틱 모형에서의 $\pi(x)$는 $X=x$일 때의 성공확률이며, 이항분포의 모수가 된다. $\pi(x)$의 함수, $\log it(\pi(x))$는 S-곡선 형태인 증가 또는 감소하는 곡선을 나타낸다. 또한 위의 식 (1)을 변환하여 식 (2)로 성공확률을 직접 표현할 수 있다.

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \qquad (2)$$

식 (1)과 같은 모형을 선형확률모형(linear probability model)이라 부르며 기울기 β는 x가 1단위 변할 때 성공확률의 변화량을 나타낸다. 또한 β의 부호로 곡선의 증가 또는 감소를 알 수 있다.

1.1. 오즈비(Odds ratio)

로지스틱 모형에서는 오즈(odds)와 오즈비(odds ratio)를 생각할 수 있다. 식 (1)에서 성공의 오즈는

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\alpha+\beta x) = e^{\alpha}(e^{\beta})^{x} \quad (3)$$

로 x가 1단위 증가함에 따라 오즈는 e^{β}배만큼 곱해져 증가함을 알 수 있다. 오즈의 범위는 음이 아닌 실수이며 실패보다 성공 가능성이 더 높은 경우 1보다 큰 값이 된다.

오즈비는 한 그룹의 사건이 발생했을 때의 오즈($odds_2$)에서 다른 그룹의 사건이 발생했을 때의 오즈($odds_1$)의 비로

$$\theta = \frac{odds_1}{odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \quad (4)$$

와 같이 나타낸다. 오즈비가 1이라는 것은 두 집단 사이의 발생할 사건의 오즈가 같음을 나타내며, 1보다 큰 경우는 그룹 1이 더 많이 일어 날 수 있음을 의미한다. 이러한 오즈비는 이항자료의 값 사이의 관련성이나 비독립성의 정도를 표현하는 척도로 로지스틱모형의 해석에 있어 예측변수의 효과를 설명할 수 있다 (Agresti, 1996).

1.2. 신뢰구간

로지스틱 모형 $\log it(\pi(x)) = \alpha+\beta x$ 에서 모수 β에 대한 대표본 신뢰구간은

$$\hat{\beta} \pm z_{\alpha/2}(ASE) \quad (5)$$

로 여기서 ASE는 asymptotic standard error를 의미한다. 이 구간을 지수 변환하면 오즈의 증가배율의 e^{β}의 신뢰구간을 얻을 수 있다.

1.3. 검정

귀무가설은 $H_0 : \beta = 0$으로 성공확률이 X와 관계가 없음을, 즉 이항반응에 미치는 X의 효과가 없음을 주장한다. 이를 검정하기 위하여 필요한 검정통계량은 대표본을 가정한 경우

$$Z = \frac{\hat{\beta}}{ASE} \tag{6}$$

와 같다. 또한 동일한 귀무가설을 검정할 수 있는 왈드통계량(Wald statistic)은 $(\frac{\hat{\beta}}{ASE})^2$로 자유도가 1인 카이제곱분포를 따른다. 대표본일 때는 왈드통계량이 좋지만, 실제로는 우도비검정이 더 우수한 방법으로써 주로 사용되고 있다 (정광모 외 1인, 2006). 우도비 검정통계량은 $-2(L_0 - L_1)$로 L_0는 $\beta = 0$일 때 즉, 귀무가설 하에서의 최대로그우도이며, L_1는 β에 대한 무제한의 가정 하에서 구해진 최대로그우도로 이 통계량도 자유도 1인 카이제곱 근사분포를 따른다.

2. 모형평가

구축된 모형을 평가하는 것은 모형을 구축하는 것 못지않게 중요한 절차이다. 모형을 평가하는 방법에는 잭나이프(Jackknife) 방법과 부스트랩(bootstrap) 방법을 이용한 재표본(resampling) 방법 등이 있지만, 본 연구에서는 교차타당성 방법을 소개하고자 한다. 또한, 교차타당성 방법을 사용하여 얻은 예측값으로 ROC 곡선을 그려 곡선의 면적과 신뢰구간(confidence limit)을 구하여 ROC 곡선을 평가하고자 한다.

2.1. 교차타당성 (Cross-validation)

데이터가 주어졌을 때, 연구자는 우선 데이터의 구조를 적절히 설명할 수 있는 모형을 구축하고자 할 것이다. 모형을 구축한 후에는 그 모형이 데이터를 얼마나 정확하게 설명하고

있는지 확인하고자 할 것이다. 연구자는 모형을 구축하는데 주어진 데이터를 사용하였으므로, 데이터에 적합하다고 간주되는 모형을 평가하기 위해서는 사용된 데이터와는 독립적인 데이터가 있어야 한다. 그러나 모형 평가를 위해서 독립적인 데이터를 이용하는 것은 현실적으로 매우 어려운 일이다. 이러한 문제점을 해결하기 위해 모형을 예측하고 평가하는데 교차타당성을 이용한다.

교차타당성은 주어진 데이터를 가지고 모형의 예측과 평가를 동시에 할 수 있는 방법을 제공하고 있다. 교차타당성 방법을 이용하기 위해서는 가지고 있는 데이터를 트레이닝 데이터(training set)와 테스팅 데이터(testing set)로 구분하여야 한다. 여기서 트레이닝 데이터는 모형을 구축하는데 쓰이는 데이터를, 테스팅 데이터는 모형을 평가하는데 쓰이는 데이터를 말한다(http://en.wikipedia.org/wiki/Cross-validation_(statistics)).

2.1.1. Leave-one-out approach

Leave-one-out 교차타당성 방법은 주어진 데이터에서 하나의 자료만을 제외한 나머지 데이터를 트레이닝 데이터로 제외된 하나의 자료를 테스팅 데이터로 사용한다. 그러므로 이 방법을 사용한다면 주어진 데이터의 각각의 자료가 한 번씩은 적합된 모형을 평가하는데 사용되게 되는 것이다. 이 같은 방법은 각각의 테스팅 데이터가 상호 배타적이라는 것과 전체 데이터를 효과적으로 이용한다는 것이 장점이 될 수 있는 반면, 전체 데이터의 자료의 수만큼 모형을 예측하고 평가하는 과정이 필요하게 되기 때문에 비효율적이라는 단점이 있다(Pang-Ning Tan et al., 2006). Leave-one-out 교차타당성 방법에서 average error rate는 전체 자료의 수를 N이라 할 때 $\frac{1}{N}\sum_{i=1}^{N} Error_i$와 같이 나타내어진다.

2.1.2. K-fold approach

K-fold 교차타당성 방법은 주어진 데이터를 K개의 부표본(subsample)으로 나누어 $K-1$개의 부표본을 트레이닝 데이터로, 나머지 1개의 부표본을 테스팅 데이터로 이용하게 된다. 이때 전체 데이터를 나눌 때는 부표본의 크기가 모두 동일하게끔 $\frac{1}{K}$로 나누는 것이 일반적이다. 예를 들어 3-fold 교차타당성 방법을 생각해보자. 전체 데이터를 $\frac{1}{3}$씩 나누어 3개의 부표본을 만들어 2개의 부표본은 모형을 적합시키고, 나머지 1개의 부표본은 모형을 평가한다. 이와 같은 절차는 개개의 부표본이 모형을 평가하는데 이용되게끔 총 3번을 반복하게 된다. 이는 일반적인 K-fold 교차타당성 방법에서와 동일하게 적용된다. 그러므로 K

-fold 교차타당성 방법에서도 마찬가지로 개개의 자료마다 오직 한 번씩만이 테스팅 데이터로 사용된다 (Pang-Ning Tan et al., 2006). K-fold 교차타당성 방법에서 average error rate는 $\frac{1}{K}\sum_{i=1}^{K} Error_i$이다.

2.2. ROC 곡선 (The Receiver Operating Characteristic Curve)

교차타당성 방법은 모형을 평가하는데 이용함으로써 적합된 모형의 예측정도를 추정할 수 있다는 것이 하나의 장점이다. 선형 모형에서는 MSE(Mean Squared Error), MAD(Median Absolute Deviation) 등의 척도를 사용한다. 그러나 로지스틱과 같은 선형모형에서 MSE, MAD는 간단한 공식으로 해결되지 않는다. 그러므로 본 연구에서는 ROC 곡선을 이용하도록 한다. Leave-one-out과 K-fold 교차타당성 방법을 사용했을 때의 예측정도를 ROC 곡선으로 표현할 수 있는 것은 두 가지 방법 모두 전체 데이터의 개별 자료가 오직 한 번씩 모형을 평가하는데 사용되었기 때문이다 (Gönen, 2007).

ROC 곡선은 각각의 점에서 cut-off value에 따라 TPR(True Positive Rate)을 y축에, FPR(False Positive Rate)을 x축에 표시함으로써 그려진다. [표 1]에서 처럼 $TPR = \frac{TP}{(TP+FN)}$, $FPR = \frac{FP}{(TN+FP)}$이다 (Pang-Ning Tan et al., 2006).

ROC 곡선 아래의 면적을 AUC(the Area Under the ROC Curve)라 한다. AUC는 구축된 모형을 평가하는 하나의 방법으로 만약, 모형이 완벽하게 data에 적합되었다면 AUC는 1이 될 것이다.

<표 1> 분류 행렬표

		Predicted Class	
		+	-
Actual Class	+	True Positive(TP)	False Negative(FN)
	-	False Positive(FP)	True Negative(TN)

3. 장기손해보험

손해보험은 보험자가 우연한 사고(보험사고)로 생기는 손해를 전보(塡補)할 것을 약정하

고, 보험계약자가 이에 보험료를 지불할 것을 약정하는 것으로서 보상내용, 법률에 의한 강제 유무, 보험 목적, 마케팅 관점, 보험금액 결정방법 등 그 분류 방법에 따라 여러 가지 형태로 나눌 수 있다 (손해보험협회, http://www.knia.or.kr). 장기손해보험은 자동차보험, 화재보험, 특종보험, 해상보험, 개인연금손해보험등과 같이 보상내용에 따른 손해보험 종류에 속한다.

장기손해보험은 통상 손해보험상품의 보험기간이 1년인데 반해 보험기간이 3년 이상인 손해보험을 말한다. 손해보험사에서 판매하는 건강보험, 어린이보험, 암보험, 간병보험, 통합보험, 장기화재보험이 모두 장기손해보험에 해당된다. 장기손해보험은 상해, 질병, 화재, 배상책임을 보장하며 매월 일정금액의 보험료를 만기까지 납입하고 납입보험료 중 일부를 만기에 환급 받는다. 장기손해보험의 특성을 살펴보면, 보험기간이 3년 이상으로 장기간이며, 예기치 못한 불의의 사고에 대비할 수 있는 보장기능과 함께 만기시에는 계약자가 납입한 보험료 중 저축보험료 부분에 약정된 예정이율에 따른 이자를 더해 돌려주는 저축기능을 겸한 보험상품이다 (박상범 외 2인, 2006). 보험상품은 제1회 보험료 납입일 16시부터(단, 암 관련 보장상품은 제외) 보험계약의 효력이 발생되어 약정기간동안 보험계약의 효력이 지속되며 동일상품의 판매를 중지한다 하더라도 기존 가입자의 보험계약은 소멸되지 않고 보험기간 만료시까지 효력이 계속된다. 계약의 당사자는 보험계약자, 피보험자, 사망보험금수익자, 보험회사로 구성된다. 보험계약자는 보험계약을 체결하고 보험료 납입의무를 지는 자이며, 피보험자는 보험사고 발생의 대상이 되는 자이다. 사망보험금수익자는 보험사고 발생시 사망보험금 청구원을 가진자이며, 보험회사는 보험금 지급 의무를 지는 보험회사를 말한다. 보험가입시 나이는 만나이+6개월을 한 나이로 이를 보험나이라고 지칭한다. 만기환금급을 비율로 표현한 만기환금급율은 환급금/누적보험료로 계산된 값이다.

4. 데이터 적용

A 보험회사의 장기손해보험 상품에 가입한 보험자들의 자료를 이용하여 로지스틱 회귀모형을 적합하고, 교차타당성 방법을 사용하여 모형을 평가해 보고자 한다. 본 연구에서 사용된 자료는 2003년 11월 17일에서 19일 사이에 보험이 개시된 사람들로 보험계약자와 피보험자가 동일한 178명을 대상으로 2008년 5월 당시 보험을 유지, 해지 또는 실효 상태 및 보험 관련 정보가 명시되어 있다. 연구에서 사용된 변수와 자료의 형태는 [표 2]와 같다.

[표 2]의 자료에서 볼 수 있는 변수들의 특징을 살펴보면, 총 변수의 수는 종속변수인 보험유지상태 변수를 포함하여 9개이고, 보험유지상태는 이항자료로 0인 경우가 해지, 1인 경

우가 유지임을 나타낸다. 본 연구에서는 보험 계약이 부활할 수 있는 실효는 유지의 경우로 포함시켜 분석하였다. 보험해지나이는 보험을 해지한 경우에는 해지당시의 보험나이, 보험을 유지하고 있는 경우에는 실제 현재의 보험 나이를 가리킨다.

<표 2> 장기손해보험 데이터

id	보험 유지상태	성별	가입나이	보험 유지개월	보험 만기년수	만기환급 금율	보험납부 년수	만기년수-납수변수	보험해지 나이
1	0	남	43	33	10	106	10	0	45
2	0	남	44	34	10	115	5	5	46
3	0	여	30	48	10	104	7	3	34
4	0	여	26	3	10	103	7	3	26
5	0	남	40	6	10	101	7	3	40
6	1	여	41	54	10	103	7	3	45
7	1	여	36	54	10	103	10	0	40
8	0	남	43	45	10	116	7	3	46
9	0	여	40	14	10	101	10	0	41
10	0	남	43	12	10	109	5	5	44
..	..	..	..	..	..	..	..	..	..

4.1. 기초 통계량

<표 3> 장기손해보험 데이터의 기초 통계량

변수		빈도(%)	중위수
성별	남	58(48.74)	
	여	61(51.26)	
가입당시나이	11~20살	7(5.88)	40
	21~31살	13(10.92)	
	31~40살	48(40.34)	
	41~50살	45(37.82)	
	51~60살	3(2.52)	
	61~70살	3(2.52)	
보험유지여부	유지, 실효	46(38.66)	
	해지	73(61.34)	
만기년수	10년	66(55.46)	10
	15년	53(44.54)	
만기환급금율	100~110(%)	89(74.79)	106

	111~120(%)	26(21.85)	
	121~130(%)	2(1.68)	
	131~140(%)	2(1.68)	
보험금납입년수	3년	18(15.13)	7
	5년	26(21.85)	
	7년	43(36.13)	
	10년	15(12.61)	
	15년	17(14.29)	
보험료	0~100000원	84(70.59)	100000
	100001~200000원	25(21.01)	
	200001~300000원	2(1.68)	
	300001~400000원	1(0.84)	
	400001~400000원	7(5.88)	

[표 3]에서는 장기손해보험 상품의 데이터에 대한 기초 통계량을 제시하였다. [표 3]을 보면 장기손해보험을 가입한 고객의 나이 연령은 대부분 31~50살로 약 78.16%가 차지하고 중위수는 40살임을 알 수 있다. 또한 보험유지여부에 관해서는 고객 중 61.34%가 보험을 해지하였고, 38.66%가 유지하고 실효한 상태였다. 만기환급금율인 경우는 고객의 74.79%가 100~110%, 21.85%가 110~120% 임을 알 수 있으며 중위수는 106%이었다. 보험료는 고객의 70.59%가 0~100000원을 납부하는 것을 쉽게 알 수 있다.

4.2. 로지스틱 회귀모형 적합

우리의 관심은 어떤 변수가 보험을 해지하는데 영향을 미치는지를 알고자 하는 것으로 이항자료인 종속변수를 고려하여 로지스틱 회귀모형을 구축하였다. 모형을 구축하는 데에는 성별, 가입나이, 보험만기년수, 만기환급금율, 보험납부년수, 만기년수-납부년수, 보험해지나이 총 7개의 설명변수들을 사용하였다. [표 2]에서 제시된 보험유지개월 변수는 보험해지나이를 계산하기 위하여 사용하였기 때문에 이 두 변수는 종속적인 관계가 있다. 이 같은 사실로 분석 시에는 보험유지개월 변수를 포함하지 않았다. 또한, 모형에서의 변수 선택방법은 backward 방법을 이용하였다. 적합된 로지스틱 회귀모형의 모수 추정치는 [표 4]에 제시하였고 모형식은 식 (3)과 같다.

<표 4> 로지스틱 회귀모형의 모수 추정치와 왈드통계량

모수	자유도	추정치	표준오차	왈드통계량	유의확률
intercept	1	2.2867	0.8032	8.1042	0.0044
만기환급금율	1	0.0661	0.0304	4.7307	0.0296
해지나이	1	-0.0679	0.0192	12.4735	0.0004

$$\log it(\pi(x)) = \log(\frac{\pi(x)}{1-\pi(x)}) = 2.2867 + 0.0661 * \text{만기환급금율} - 0.0679 * \text{해지나이} \quad (7)$$

[표 4]에 제시되어 있는 왈드통계량을 통하여 만기환급금율과 해지나이는 장기손해보험을 해지하는 데에 영향을 미치는 설명변수임을 알 수 있다. 또한 장기손해보험을 해지하는 데에 얼마나 영향을 미치는지 살펴보기 위해서 [표 5]에 제시된 오즈비를 살펴보도록 한다.

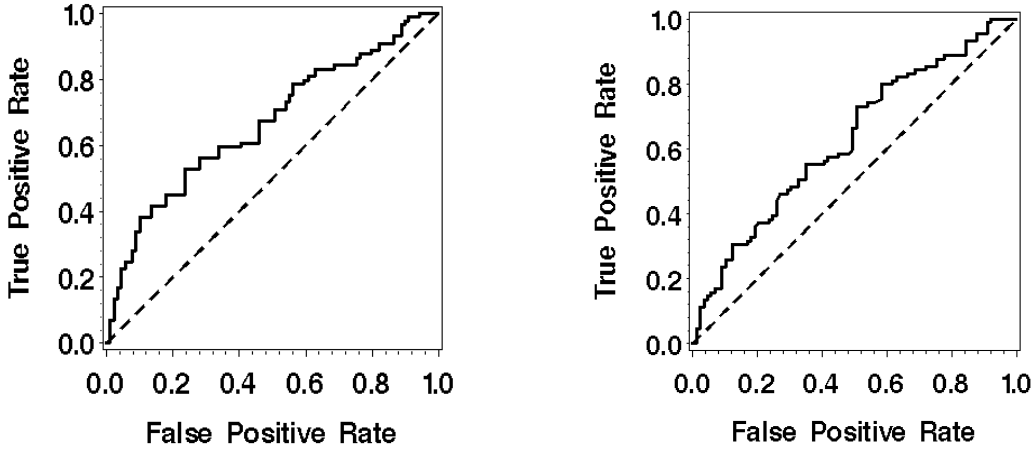
<표 5> 모수에 대한 오즈비 추정치와 95% 왈드 신뢰구간

모수	오즈비 추정치	95% 왈드 신뢰구간
만기환급금율	1.068	1.007, 1.134
해지나이	0.934	0.900, 0.970

만기환급금율인 경우는 오즈비의 추정치가 1.068로 만기환급금율이 1단위 증가 할 때 보험해지율이 1.068배 증가함을 알 수 있으며 95% 왈드 신뢰구간은 (1.007, 1.134)이다. 이와 더불어 해지나이는 오즈비 추정치는 0.934로 해지나이가 1단위 증가할 때 해지율이 0.934배증가함을 알 수 있으며 95% 왈드 신뢰구간은 (0.900, 0.970)로 보험을 가입한지 오래 될수록 보험해지율이 낮아짐을 알 수 있다.

4.3. 교차타당성 방법에 따른 ROC 곡선

앞에서 구축한 로지스틱 회귀모형을 평가하기 위하여 leave-one-out 교차타당성과 10-fold 교차타당성 방법을 사용하였다. 이 두 가지 방법을 이용하여 장기손해보험의 해지율을 예측하여 본 결과를 ROC 곡선으로 나타낸 것이 [그림 1]이다. [그림 1]의 ROC 곡선을 보면 leave-one-out 교차타당성과 10-fold 교차타당성 방법의 예측정도가 많은 차이를 보이지 않고 있다고 할 수 있다. 이는 또한 [표 6]에 제시되어 있는 AUC를 통하여



(a) leave-one-out cross validation (b) 10-fold cross validation

[그림 1]. Leave-one-out과 10-fold 교차타당성 방법을 사용한 ROC 곡선

leave-one-out과 10-fold 교차타당성 각각 0.6660과 0.6288로 비슷하다는 것을 알 수 있다. 또한 이들의 95% 신뢰구간을 살펴보면 leave-one-out 방법은 (0.5864, 0.7455)이고, 10-fold 방법은 (0.5472, 0.7105)으로 앞의 로지스틱 회귀모형의 장기손해보험 해지율을 예측하는데 있어서 월등히 좋지는 못하다는 것을 알 수 있다.

<표 6> 교차타당성 방법에 따른 ROC 곡선의 면적과 95% Confidence Limits

	AUC	standard error	95% Confidence Limits
leave-one-out	0.6660	0.0406	0.5864, 0.7455
10-fold	0.6288	0.0417	0.5472, 0.7105

Ⅲ. 결론 및 향후과제

본 연구에서는 장기손해보험 상품의 해지에 관하여 알아보기 위해 로지스틱 회귀모형과 교차타당성, ROC 곡선에 대하여 살펴보고 이를 실제 데이터에 적용시켜 보았다. 그 결과 장기손해보험을 유지 하는 데 영향을 미치는 변수는 만기환급금율과 해지나이임을 알 수 있었다. 만기환급금율이 증가하면 보험해지율은 증가하는 반면, 해지나이가 증가하면 보험해지율은 감소함을 오즈비를 통하여 알 수 있었다. 만기환급금율이 증가할수록 보험 해지율이 증가하는 것은 만기 환급금율이 높을수록 가입 기간내에 추가 납입하는 보험료의 상승과 더

불어 장기 상품인 경우 화폐 가치와 만기환금금율의 상대적 비율로 인해 초기에 높은 만기환급금율이 높은 상품을 가입이 경우 해지율에 영향을 주는 것으로 알려져 있다. 교차타당성을 이용이 ROC곡선으로 로지스틱 회귀모형을 평가하였다. leave-one-out 교차타당성 방법과 10-fold 교차타당성 방법을 이용하여 적합된 모형이 보험해지율을 정확하게 예측한다고 할 수 없음을 알 수 있었다. 이는 보험해지에 외부효과 및 여러 요인 등이 영향을 주었음을 생각해 볼 수 있다. 또한, 본 연구에서는 보험계약자와 피보험자가 동일한 사람들의 보험 가입사항으로 변수 7개만으로 장기보험험상품의 해지에 미치는 영향력을 연구하는 한계가 있다. 그러므로 장기손해보험 해지율을 추정하는데 있어 이에 영향을 주는 효율적인 요인을 찾기 위하여 보험 가입자에 대한 구체적이고 체계적인 정보의 수집이 필요하다. 향후에는 다양한 계층의 보험 대상자와 여러 장기보험상품의 정보들을 통해서 보험해지에 영향을 미치는 공통된 요인을 찾아보는 것이 필요하다고 생각되어진다.

참고문헌

[1] 박상범, 김용하, 이희춘 , 2006. 손해보험론. 문영사.

[2] 염준근, 2005. 선형회귀분석. 자유아카데미.

[3] 이경룡, 2002. 보험학원론. 영지문화사

[4] 정광모, 최용석, 2006. 범주형 자료분석 개론 -SAS의 응용 및 해석-. 자유아카데미.

[5] 허연, 2006. 생활과 보험. 문영사

[6] Agresti, A. 1996. *An Introduction to categorical data analysis.* John Wiley & Sons, INC.

[7] Gönen, M.. (2007). *Analyzing receiver operating characteristic curves with SAS.* SAS Press Series.

[8] Tan, Pang-Ning, Steinbach, M., and Kumar, V. (2006). *Introduction to data mining.* Pearson Education.

[9] http://en.wikipedia.org/wiki/Cross-validation_(statistics)

[10] 손해보험협회 http://www.knia.or.kr

[투고일: 2009. 08. 10][심사(수정)일: 2009. 08. 15][게제확정일: 2009. 08. 19]

Annals of Statistical Research
Vol. 6. No. 1, 2009, pp.41~55

Won / dollar exchange rate using time series data, research on prediction models

Sung-Jin Ahn[1], Young-Jin Jung[2], In-Sup Kim[2]

Abstract

In this paper, pursuant to a change in the institutionalized since the financial crisis in the won / dollar in the average monthly pursuant to volatility through many time series forecasting models and forecasting models for each time series was performed to compare the predictions.

Key Words: time series, comparison, volatility, forecating, ARIMA, Transfer function model, exponential smothing

1) Corresponding Author : Graduate student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : bmyang28@naver.com
2) Graduate student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : jungjin0124@hanmail.net
2) Graduate student, Department of Statistics, Dongguk University, Pildong 3-26, Joonggu, Seoul, Korea. 100-715, E-mail : supik031@nate.com

Ⅰ. 서 론

1. 연구의 개요

환율제도의 유형에는 환율을 일정수준으로 고정시키거나 환율 변동 폭을 일정 수준으로 정하여 놓고 이러한 환율의 유지를 위해 정책당국이 외환시장에 개입하는 고정환율제와 외환시장의 수요와 공급에 의해 자율적으로 결정되도록 하고 정책당국이 외환시장에 개입하지 않는 자유변동환율제가 있다. 우리나라의 환율 제도를 보면 IMF를 기점으로 전은 고정환율제를 후로는 자유변동환율제로 바뀌게 되었으며 이후 환율 예측은 많은 관심과 연구의 대상이 되어 왔다. 외환위기 이후 외환자유화 및 자본자유화가 확대되고 개방화 및 국제화의 진전, 무역규모의 지속적인 증가로 인해 외화자산 및 부채의 규모가 커짐에 따라 외환 거래량은 늘어나고 외환관련 금융상품도 한층 더 활발히 거래되고 있다. 정부는 환율의 변동성이 지나치게 큰 경우를 제외하고는 외환시장에 거의 개입하지 않기 때문에 환율의 예측이 이전보다 더 어려워진 측면이 있다.

본 연구에서는 외환위기 이후의 자유변동환율 제도하에서의 원/달러 평균 월별 환율의 변동률을 여러 시계열 예측 모형을 통하여 예측하고 각 시계열 모형의 예측력의 비교를 수행하였다. 아울러 자유변동환율 제도 하에 원/달러 환율을 예측하고자 하는 기관에 시계열 분석방법의 의한 예측 모형에 대한 기준을 제시해 줄 수 있다는 차원에서 의의를 찾을 수 있을 것이라고 본다.

연구의 순서는 원/달러 월 평균 환율 데이터를 이용하여 원/달러 월 평균 환율의 변동률을 구하였다. 그리고 여러 시계열 예측 모형을 추정한 다음 각 시계열 예측 모형을 이용하여 예측을 하였다. 각 모형에 의해 예측된 값과 실제 값을 이용하여 예측오차를 구하여 각 모형의 예측력을 평가하였다. 기존의 원/달러 환율 변동률에 대한 연구에서 추정한 여러 예측 모형을 임의확률(Random Walk) 모형과 비교하여 모형을 평가하였다. 본 연구에서도 같은 방법을 이용하여 각 시계열 예측 모형에 대한 예측력을 평가하였다.

2. 기존연구

환율 예측과 관련된 국내 문헌은 1990년 중반 이후 많이 나오기 시작했다. 초장기 환율 예측에 관한 연구들은 주로 우리나라 환율결정에 있어 중요한 변수가 무엇인지에 대한 분석에서 출발하였다. 이후 이종욱(1992)은 1980년 1/4분기부터 1990년 3/4분기의 자료를 이

용하여 Hooper and Morton(1982)모형과 Yoshikawa(1990)에 의한 공급중심 환율결정 모형을 변형시킨 모형을 바탕으로 환율예측을 시도하였지만 표본내 예측만 수행하였다.

이근영(1997)은 1985년 4월부터 1996년 5월까지의 월별 원/엔 재정환율을 이용하여 다양한 예측기간을 놓고 예측력을 비교하였는데 단가예측에는 임의보행 모형의 예측력이 저조하긴 하나 예측력에 큰 차이를 보이지 않은 반면 장기예측의 경우에는 AR모형이 임의보행 모형보다 유의적으로 높은 예측력을 보인다는 결과를 보고하였다.

전선애(1998)는 1980년 1/4분기부터 1996년 4/4분기의 자료를 이용하여 7가지 분기별 원/달러 예측모형을 비교하였다. 예측기간이 1년 미만인 단기예측에 있어서는 대부분의 모형에서 오차수정을 반영한 모형이 임의보행 모형보다 예측력이 높음을 보여준 반면 예측기간이 1년을 넘는 장기예측에서는 차분방정식 모형이 보다 우수한 예측력을 가지는 것으로 나타났다.

오문석·이상근(2000)의 경우 1990년 3월부턴 1996년 12월까지의 월별자료를 이용하여 자산시장접근모형, 임의보행모형, ARIMA 모형의 장단기 예측력을 비교하였는데 표본외 예측에서는 1, 3개월의 초단기에는 임의보행모형이 , 6, 12개월의 예측결과에서는 실질이자율 차모형과 종합모형의 예측력이 우수한 것으로 나타났다. 한편 표본내 예측력에서는 모든 기간에서 임의보행 모형이 가장 우수한 것으로 나타났다.

정철호(2004)는 1997년 8월부터 2002년 2월까지의 주별 데이터를 이용하여 표본외 예측을 한 결과 선물환 환율 정보를 포함한 벡터오차수정모형의 예측력이 전반적으로 임의보행 모형을 능가하는 것으로 나타났다. 한편 신동백(2006)의 경우 단순회귀 분석을 통해 여러 경제 변수 중에서 환율에 미치는 영향이 높은 변수들로 구성된 모형으로 예측치를 구한 다음 이를 실제 환율과 비교하였다.

이상에서 살펴본 바와 같이 원/달러 환율 예측에 대한 연구결과들은 임의보행 모형과의 예측력을 비교를 바탕으로 단순한 시계열 모형이나 경제이론에 입각한 모형들을 제시하고 있다. 대체적인 결론은 단기예측에 있어서는 임의보행 모형의 예측력이 우수하나 장기예측에 있어서는 경제이론에 입각한 구조모형의 예측력이 우수한 것으로 나타났다(이윤석, 2007).

최근에 들어서는 비선형성을 강조한 복잡한 형태의 모형에 대한 연구들이 많아지고 있다. 김봉함·유만식(2004)은 다양한 형태의 마르코프 국면전환 모형을 설정하고 전이확률이 고정된 Hamilton 모형을 사용하였으나 임의보행 모형보다 정확한 예측을 하지 못하였다.

박범조(1997)는 신경회로망 회귀위수 모형을 바탕으로 4개 통화의 외한수익률 예측을 시도하고 표준신경회로망 모형이나 ARMA 모형보다도 우수한 예측력을 보고하였다. 신성환

(1995)도 신경회로망 모형을 이용하여 분석을 하였으나 노드 수에 따른 예측력 차이를 설명하지 않고 있어 다소 아쉬운 점으로 지적되었다.

또한 환율 자체의 예측보다는 환율의 변동성에 대한 예측력을 비교한 연구결과들도 발표되고 있다. 최생림·형남원(2003)은 지수적 이동가중평균(EWMA: exponentially weighted moving average) 방식을 이용하여 예측한 결과 단기에서는 GARCH유형의 시계열 모형보다 상대적으로 예측력이 우수하나 1개월 이상의 장기에서는 FIGARCH모형의 예측력이 뛰어나다는 결과를 보고하였다.

Ⅱ. 본론

1. 데이터

본 논문에서는 한국은행 경제통계시스템에서 제공하는 원/달러 월 평균 환율 데이터를 이용하였다.[3] 원/달러 월 평균 환률 데이터를 이용하는 것은 대부분의 국가에서 환율에 대한 기준으로 사용하고 있기 때문이다. 본 논문에서 원/달러 월 평균 환율의 자료는 1998년 4월부터 2009년 7월까지 총 136개의 관측값을 사용하였다.

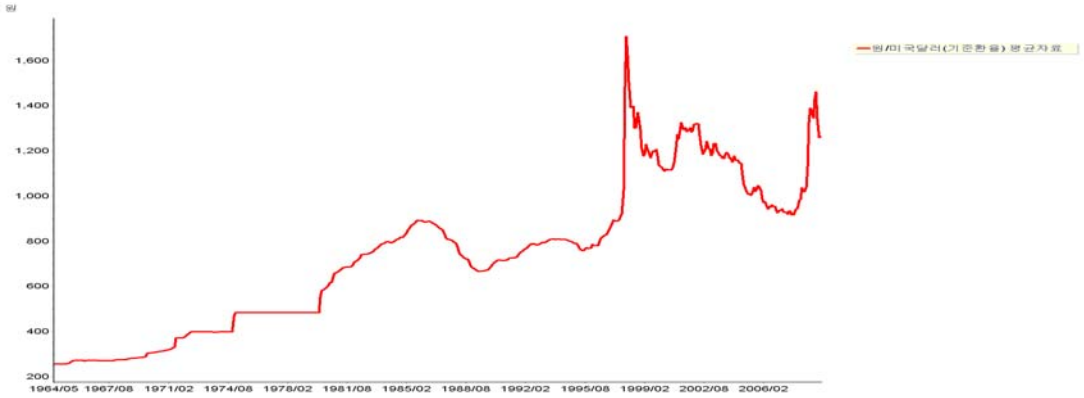
[그림 1]는 1964년 5월부터 2009년 7월까지 나타낸 원/달러 월 평균 환율을 나타낸 시계열 자료의 시도표이다. 1997년도 12월부터 1998년 3월까지는 외환위기에 따른 영향을 확인 할 수 있다. 본 논문에서는 1998년 3월까지 외환위기에 영향을 받은 것으로 판단하여 1998년 4월의 원/달러 월 평균 환율을 이용하여 예측모형을 구축하여 각 모형들의 예측력을 비교하였다.

원/달러 월 평균 환율의 변동률은 원/달러 월 평균 환율을 로그 차분하여 얻었다.

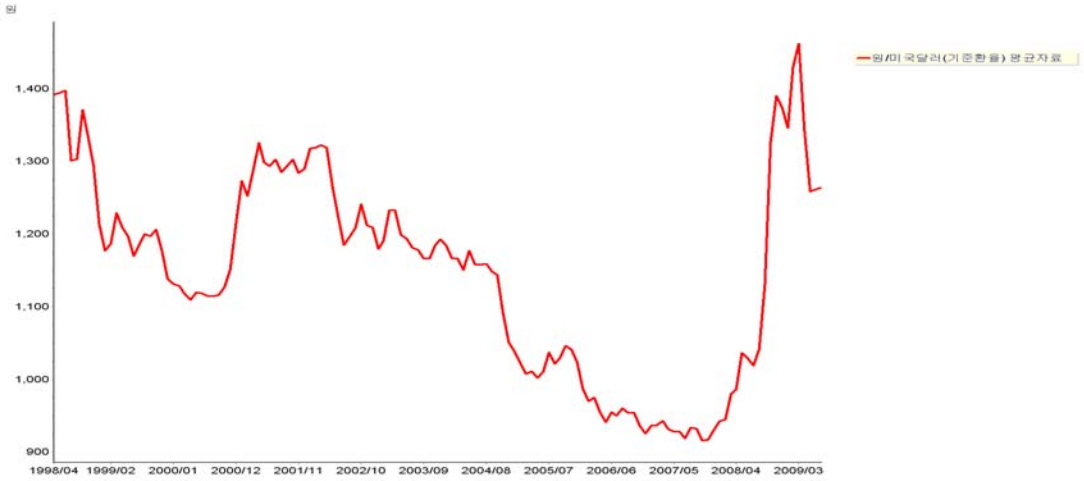
$$Rate_t = \log\left(\frac{extr_t}{extr_{t-1}}\right) \times 100$$

여기서 $Rate_t$는 t시점에서의 원/달러 월 평균 환율의 변동률을, $extr_t$은 t시점에서의 원/달러 월 평균 환율을 나타낸다.

3) 한국은행 경제통계시스템 http://ecos.bok.or.kr/



[그림 2] 원/달러 월 평균 환율(1964년 5월~2009년 7월)



[그림 3] 원/달러 월 평균 환율(1998년 4월~2009년 7월)

[그림 3]는 이 논문에 이용된 기간의 원/달러 월 평균 환율을 나타낸 표이다.

2. 원/환율 월 평균 환율의 변동율의 특성

각 시계열 예측 모형의 추정을 위해 우선 1998년 5월부터 2009년 6월까지의 원/달러 환율의 변동률 데이터의 분포를 살펴보았다. [표 1]에는 월별 월/달러 환율의 변동률의 관측치의 수, 평균, 표준편차, 최소값, 최대값, 왜도, 첨도 그리고 Jarque-Bera(JB) 통계량을 정리한 결과이다.

<표 1> 원/달러 환율 변동률에 대한 기초 통계량

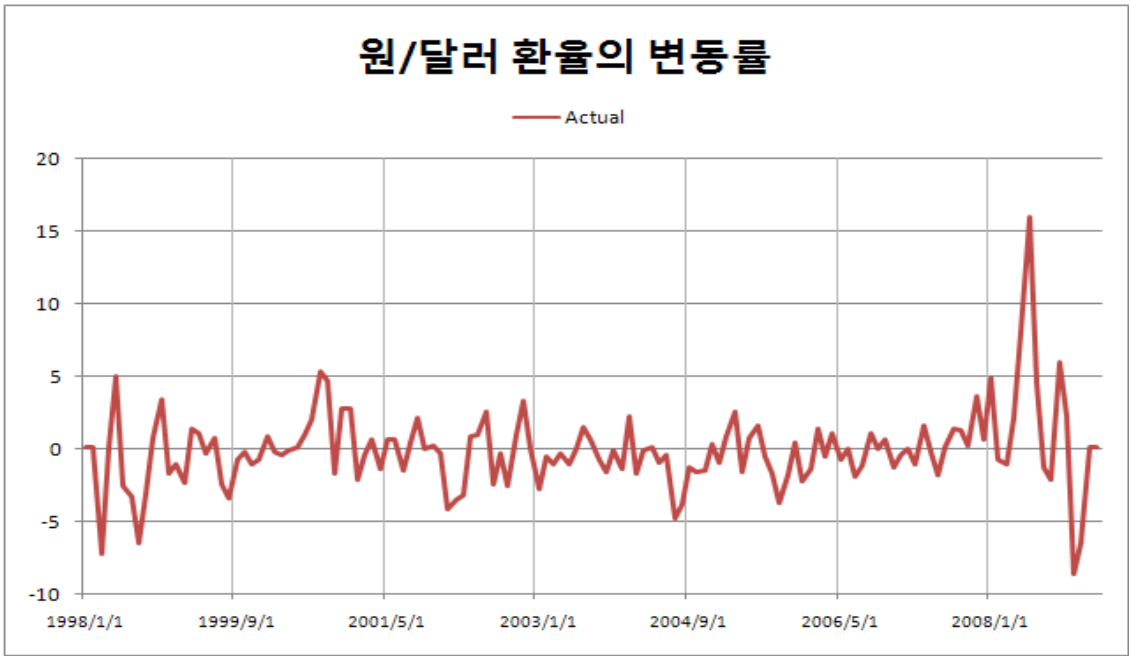
통계량	원/달러월 평균 환율의 변동률
관측치 수(Observation Number)	122
평균(Mean)	-0.2474
표준편차(Standard Deviation)	2.0642
최소값(Minimum)	-7.1500
최대값(Maximum)	5.3489
대칭도(Skewness)	-0.0889
첨도(Kurtosis)	4.4713
Jarque-Bera(JB)	11.1650

대칭도는 평균 근처의 비대칭도를 나타내는 것으로 정규분포를 따르면 0에 가까운 값을 가진다. [표 1]에서 1.358의 양의 값을 가지므로 꼬리가 양의 방향으로 위치하는 비대칭적인 분포라는 것을 알 수 있다. 다음으로 첨도는 정규분포에 대해 상대적으로 솟아오른 정도(peakness)와 편평도(flatness)를 측정하는 것으로 보통의 분석 프로그램에서 표준 정규분포의 첨도는 3을 기준으로 3보다 크면 정규분포에 비해 꼬리가 두터운 분포임을 의미하고, 3보다 적으면 정규분포에 비하여 좁은 영역에 분포하고 있음을 나타낸다. 단, 본 논문에 사용된 SAS 프로그램에서는 0의 값을 기준으로 해석하게 된다. 첨도의 값이 8.744로 0보다 큰 값을 가지므로 정규분포에 비해 꼬리가 두터운 분포임을 알 수 있다. Jarque-Bara는 정규성 검정(normality test)을 위한 통계량으로 다음과 같이 정의된다(박유성 외 1인, 2008).

$$JB = n\left[\frac{S^2}{6} + \frac{(k-3)^2}{24}\right]$$

여기서 n은 관측치의 수, S는 왜도 그리고 K는 첨도를 나타낸다.

Jarque-Bara 통계량은 실증 분포가 정규분포를 따른다는 귀무가설 하에 자유도가 2인 χ^2 분포를 따른다. 99% 신뢰수준에서 임계치가 9.21이므로 원/달려 월 평균 환율의 변동성의 분포는 정규성을 따르지 않는 다는 것을 알 수 있다.



[그림 4] 원/달러 월 평균 환율 변동률

[그림 4]는 원/달러 월 평균 환율 변동률의 시계열 자료의 시도표이다.

3. 시계열 예측 모형 구축

3.1 지수평활법

지수평활법(exponential smoothing)은 시계열의 구성요소가 시간에 따라 느리게 변동하거나, 또는 변동이 느리지 않더라도 매우 규칙적인 형태를 보여주는 경우 시계열 예측 방법이다. 그리고 복잡한 이론적 배경이 없으며 쉽고 빠른 방법으로 실무적인 활용가치가 높다. 지수평활법의 기본 생각은 '최근 값에 많은 가중치를 주고 과거 값에는 적은 가중치를 주는 것'이다(박유성 외 1인, 2008). 지수 평활법은 크게 단순지수평활법(Simple exponential smoothing), 일모수 이중 지수 평활법(one-parameter double exponential smoothing), 홀-윈터스이중지수평활법(Holt-Winters' double exponential smoothing), 가법윈터스방법(additive Winters' method), 승법윈터스방법(multiplicative Winters' method)으로 나누어진다. 각 방법은 시계열 자료의 추세와 계정요인의 변화 형태 등에 의해서 결정되어 진다.

3.2 ARIMA

과거 실제 변동성의 시계열 자료를 이용하는 Box-Jenkins의 ARIMA모형은 시계열 모형 중 비용대비 예측 효율성이 우수하고 간단한 모형이다. 변동성의 시계열이 AR(p) 과정이나 MA(q) 과정을 따른다는 가정을 바탕으로 변동성을 예측할 수 있다. 시계열 자료의 변동성이 안정적일 때 효율적인 예측이 가능하다. 예측 모형은 Box-Jenkins의 3단계 방법에 의해 추정된다.

ARIMA(p,d,q) 모형은 다음과 같이 표기할 수 있다.

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i a_{t-i} + a_t$$

3.3 전이함수 모형

전이함수는 Box-Jenkins에 의하여 고안된 예측모형으로 한 시계열의 미래 값을 예측하는데 자신의 과거나 현재의 값은 물론이고 이 시계열과 인과관계를 갖고 선도하는 다른 시계열의 과거나 현재 그리고 미래의 예측 값까지도 예측에 이용함으로써 예측 효과를 높이는 모형이다. 실제로 전이함수 모형은 비교적 적은 개수의 모수를 갖는 선형시스템의 모형으로 예측오차를 크게 줄이는 모형으로 알려져 있다.

전이함수는 전이함수잡음모형의 설정, 모형검진, 예측 등의 순서로 추정되고 예측된다. 전이함수 모형은 다음과 같이 표기할 수 있다.

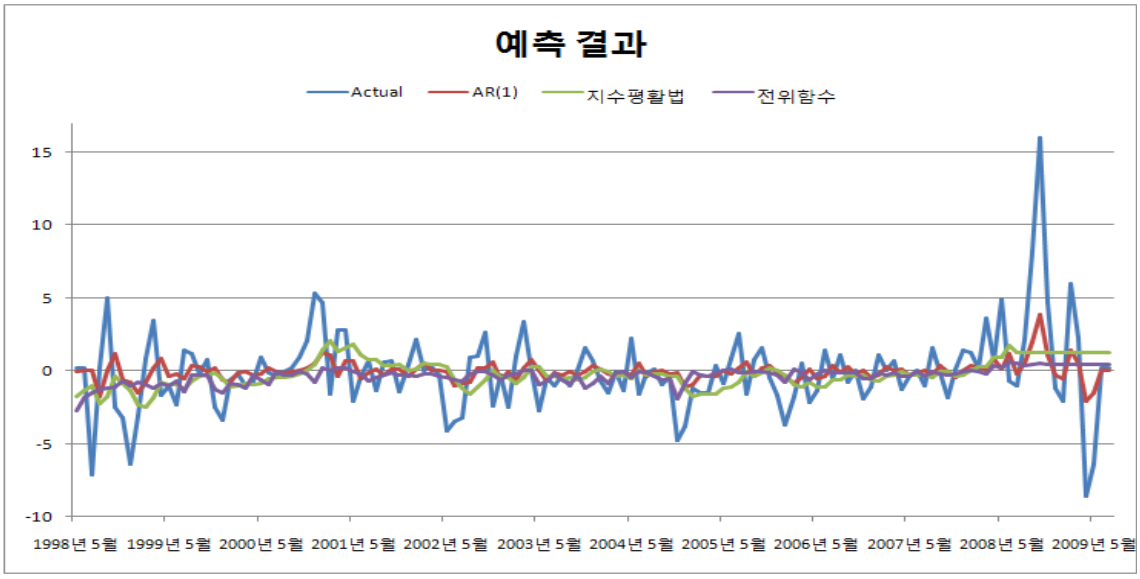
$$Y_t = v(B)X_t + N_t = \frac{w_{s(B)}}{\delta_r(B)} B^d X_t + \frac{\theta(B)\Theta(B^s)}{\phi(B)\Phi(B^s)} a_t$$

여기서 $v(B)$는 전이함수 그리고 a_t는 백색잡음 과정을 나타낸다.

3.4 예측 모형 구축 결과

<표 2> 분석결과

방법	모형	모형 식
지수평활법	가법윈터스방법	.
ARMA	AR(1)	$\widehat{Rate}_t = 0.2396\widehat{Rate}_{t-1}$
전이함수	외환 보유액 변동율	$Y_t = -0.24809X_t + \frac{1}{(1-0.61011B-0.38989B^3)}a_t$



[그림 5] 예측모형에 대한 예측 결과

[표 2]는 각 예측모형 별로 결과를 정리한 표이며 [그림 5]는 예측된 수치를 그래프로 표현한 것이다.

지수평활법은 시계열 자료의 추세와 계정요인의 변화 형태 등에 의해서 결정되어진다. 원/달러 월 평균 환율의 변동율의 시도표를 고려하여 가법윈터스방법(additive Winters' method)을 선택하여 분석하였다. ARIMA 모형은 AR(1) 모형과 MA(1) 모형 그리고 ARMA(1,1) 모형이 후보 모형으로 추정 되었고 그 중에서 예측 오차가 가장 적으면서 추정된 모수가 모두 유의한 AR(1) 모형이 최종 모형으로 선택되었다. 전환모형의 독립변수로

외한 보유액의 변동율과 코스피의 200선물 거래지수의 변동률 그리고 주가종합지수(KOSDAQ)의 변동률을 이용하였다. 하지만 외환 보유액의 변동률을 제외한 나머지 변수들은 통계적으로 유의하지 않아서 외한 보유액의 변동률이 포함된 전환모형을 추정하였다.

그리고 오차항의 추정계열인 잔차항 $\{\epsilon_t\}$에 대한 조건부 이분산성, ARCH 효과 또는 GARCH 효과가 있는지에 대해서 분석하여 보았다. $\{\epsilon_t\}$가 조건부 이분산을 가지면 $\{\epsilon_t^2\}$이 적당한 시계열 모형을 따르게 되지만 등분산성을 만족하는 것으로 결과가 나타났기 때문에 ARCH와 GARCH 모형은 고려하지 않았다.

4. 모형 비교

본 장에서는 앞장에서 살펴본 각 시계열 예측 모형들과 회귀 예측 모형의 예측력을 비교하였으며 임의확률 모형과의 예측력을 비교하여 보았다. 표본 외 예측오차를 구하는 방법을 이용하였다. 1998년4월부터 2008년 6월까지의 데이터로 추정된 각 예측 모형을 2008년 7월부터 2008년 12월까지(6개월)의 예측오차를 통하여 각 모형들의 단기간 예측력을 비교하였으며 2008년 7월부터 2009년 7월까지(13개월)의 예측오차를 통하여 각 모형들의 장기간 예측력을 비교하였다.

4.1 예측오차

시계열 모형들과 회귀 모형의 예측력을 평가하기 위하여 4가지의 예측오차를 사용하였다. 평균오차(ME: Mean Error), 평균절대오차(MAE: Mean Absolute Error), 평균절대비율오차(MAPE: Mean Absolute Percentage Error) 그리고 제곱근평균오차제곱합(RMSE: Root Mean Squared Error)을 사용하였다.

1) 평균오차(ME)

추정 예측오차의 단순평균값이기 때문에 양(+)의 오차와 음(-)의 오차가 서로 상쇄되므로 매우 작은 값을 나타낸다. 시계열 자료의 척도에 영향을 받는 단점을 가진다.

$$ME = \frac{1}{n}\sum_{t=0}^{n}(z_t - \hat{z_t})$$

여기서 n는 관측치의 수, z_t는 실제값 그리고 $\hat{z_t}$는 예측값을 나타낸다.

2) 평균절대오차(MAE)

모든 오차에 동등한 가중치를 부여하여 정확도를 측정하는 방식으로 절대값을 사용하므로 평균오차(ME)의 단점을 보완한다. 하지만 오차의 절대평균의 크기만 측정하기 때문에 오차의 상대적 크기를 알 수 없다는 단점이 있다. 그리고 시계열 자료의 척도에 영향을 받는 단점이 있다.

$$MAE = \frac{1}{n}\sum_{t=0}^{n}|z_t - \hat{z_t}|$$

여기서 n는 관측치의 수, z_t는 실제값 그리고 $\hat{z_t}$는 예측값을 나타낸다.

3) 평균절대오차(MAPE)

오차의 정도를 백분율로 똑같이 가중하기 때문에 다른 관측 수나 측정단위를 지닌 다른 예측 모형 간에 상대적인 비교가 가능하기 때문에 널리 이용되고 있다.

$$MAPE = \frac{1}{n}\sum_{t=0}^{n}\left|\frac{z_t - \hat{z_t}}{z_t}\right| \times 100$$

여기서 n는 관측치의 수, z_t는 실제값 그리고 $\hat{z_t}$는 예측값을 나타낸다.

4) 제곱근평균오차제곱합(RMSE)

가장 널리 사용되는 예측오차이다. 하지만 평균오차(ME)와 평균절대오차(MAE)와 마찬가지로 시계열 자료의 척도에 영향을 받는다.

$$RMSE = \sqrt{\frac{1}{n-k}\sum_{t=0}^{n}(z_t - \hat{z_t})^2}\,,$$

여기서 n는 관측치의 수, k는 모수의 수, z_t는 실제값 그리고 $\hat{z_t}$는 예측값을 나타낸다.

각 예측 모형의 예측력은 각 예측오차들이 작은 값을 가지는 예측모형 일수록 예측력이 높은 모형이라고 판단하였다.

그리고 각 시계열 예측 모형을 임의보행 모형(Random Walk Model)과 비교하기 위하여 임의보행 모형의 통계량과 각 예측모형의 통계량을 비율로 나타낸 테일의 U값(TU: Theil's U)을 이용하였다.

$$Theil's\ U = \frac{\text{예측모형의 통계량}}{\text{임의보행모형의 통계량}}$$

TU값이 1보다 작으면 임의보행 모형보다 상대적으로 환율 예측력이 높다는 것이고 1보다 크면 예측력이 낮음을 의미한다.

4.2 환율 예측력 비교

[표 3]은 각 예측 모형별 단기 예측오차를 나타낸 것이다. 단기 예측오차를 보았을 때에는 ARIMA 모형 > 지수평활법 > 전이함수 모형으로 나타났다. 전이함수의 모형이 가장 예측력이 나쁘게 나온 이유는 독립변수의 선택에 대한 것이라고 판단된다.

<표 3> 단기 예측오차(2008.7~2008.12)

예측오차	지수평활법	ARIMA	전이함수
ME	3.58	3.65	4.34
MAE	5.12	4.22	5.39
MAPE	120.54	77.12	109.13
RMSE	7.62	6.39	8.08

<표 4> 단기 TU(2008.7~2008.12)

예측오차	지수평활법	ARIMA	전이함수
ME	0.75	0.77	0.91
MAE	0.96	0.79	1.01
MAPE	1.29	0.82	1.17
RMSE	0.93	0.78	0.99

[표 4]은 단기 예측에서 각 예측모형과 임의보행모형을 비교하여 예측력에 차이가 있는가 없는가를 비교한 결과이다. 결과를 보면 ARIMA 모형 > 지수평활법 > 전이함수 순으로 나타났다. 모형 중 지수평활법과 전이함수 모형이 임의보행모형보다는 예측력이 떨어지는 것으로 나타났다.

<표 5> 장기 예측오차(2008.7~2009.7)

예측오차	지수평활법	ARIMA	전이함수
ME	0.35	1.20	1.11
MAE	4.56	3.45	4.51
MAPE	171.04	76.54	109.05
RMSE	6.34	4.97	6.43

[표 5]은 각 예측 모형별 장기 예측오차를 나타낸 것이다. 장기 예측오차를 보았을 때에는 ARIMA 모형 > 지수평활법 > 전이함수 순으로 예측력이 더 우수한 것으로 나타났다.

<표 6> 장기 TU(2008.7~2009.7)

예측오차	지수평활법	ARIMA	전이함수
ME	0.17	0.57	0.53
MAE	0.96	0.72	0.94
MAPE	1.31	0.59	0.84
RMSE	0.97	0.76	0.98

[표 6]은 장기예측에서 각 예측모형과 임의보행모형을 비교하여 예측력에 차이가 있는가 없는가를 비교한 결과이다. 결과를 보면 ARIMA 모형 > 지수평활법 > 전이함수 순으로 나타났다.

Ⅲ. 결론

원/달러 월 평균 환율의 변동률을 예측하기 위하여 여러 가지 시계열 예측모형을 추정하여 보았다. 추정한 결과 각 시계열 예측 모형들의 예측력은 ARIMA 모형 > 지수평활법 > 전이함수 모형 순으로 나타났다. 그리고 임의보행모형과 비교한 결과 ARIMA 모형은 단기 예측과 장기 모형에서 모두 임의보행모형보다 예측력이 우수한 것으로 나타났다. 외한위기

이후의 정부의 개입이 적은 변동환율제가 적용된 시기의 예측에는 ARIMA 모형이 가장 적합한 것으로 판단된다.

전이함수 모형의 예측력도 뛰어난 것으로 알려졌지만 이 논문에서는 독립변수의 선택에 문제가 있어서 예측력이 ARIMA 모형보다 낮게 나타났다. 따라서 여러 경제적 상황이 복합적으로 영향을 미쳐 원/달러의 환율이 결정되는 만큼 독립변수 선택에 대한 연구가 필요하다고 생각된다.

향후 과제로는 첫째, 선형적인 관계와 비선형적인 관계를 고려하여 시계열 인공신경망모형 등을 이용하여 예측하는 방법이다. 원/달러 월 평균 환율 변동률은 선형적인 요인과 비선형적인 요인이 혼재하므로 비선형적인 것과 선형적인 것의 패턴을 읽어 예측 모형에 포함시켜 미래를 예측하면 예측력이 더 높아질 것이라고 판단된다.

둘째, 다른 경제적 이론과 상황에 맞는 변수를 추가하여 시계열 회귀분석, ECM 예측모형, ADL 모형 그리고 VAR 모형 등을 이용하여 예측하는 방법이다. 경제적인 여러 상황들이 종합적으로 영향을 주어 환율이 결정되기 때문에 현재까지 고려된 여러 경제적 이론과 그리고 환율 결정에 고려가 되는 또 다른 변수들을 고려한 예측모형을 이용하면 예측력이 더 높아질 것이라고 판단된다. 그리고 마지막으로는 데이터의 기간을 달리하여 각 기간에 대하여 각 예측 방법들의 예측력을 비교하는 것도 의의가 있으리라고 판단된다.

참고문헌

[1] 김명직, 장국형(2008), 금융시계열분석 제2판, 경문사.

[2] 김봉한·유만식(2004), 마코프 국면전환모형을 이용한 환율 예측분석, 경제논집 제 43권 1-2호, pp.269~286.

[3] 김해경, 이명숙(2005), 경제 및 금융자료를 위한 시계열 분석, 경문사.

[4] 박범조(1997), 신경회로망 회귀위수를 이용한 환율 예측, 경제학 연구 제 45권 2호.

[5] 박유성, 김기환(2008), SAS/ETS를 이용한 시계열자료분석Ⅰ, 자유아카데미.

[6] 신동백(2006), 원/달러 환율 예측을 이용한 환위험분석, 산업경제연구 제19권 2호, 한국산업경제학회, pp.565~584.

[7] 신성환(2004), 마코프 국면전환모형을 이용한 환율 예측분석, 경제논집, 제 43권 1-2호, pp.269~286.

[8] 오문석·이상근(2000), 환율결정 모형의 원/달러환율 예측력 비교, 경영학연구 제29권 4호, pp.711~722.

[9] 이근영(1997), 원/엔 환율 예측모형에 관한 연구, 국제경제연구 제 3권 3호, pp.109~127.

[10] 이윤석(2007), 원/달러 환율 예측력 분석에 관한 연구, 한국금융연구원.

[11] 이종욱(1992), 총수요, 총공급과 환율결정 모형: 원화대 미환율의 실증적 연구, 금융연구, 한국금융연구원.

[12] 이종협(2007), 시계열 분석과 응용, 자유아카데미.

[13] 전선애(1998), 원/엔 환율 예측모형 개발, 환은경제연구소.

[14] 정철호(2004), 선물환 환율을 이용한 원/달러 환율 예측: VECM 기법을 중심으로, POSRI 경영연구 제 4권 1호, 포스코경영연구소, pp.174~190.

[15] 최생림·형남원(2003), 경제모형과 임의보행(Random Walk)모형의 단기 환율 예측: 엔-달러 환율의 경우, 금융학회지 제 2권 1호, pp.85~113.

[16] 조선화(2009), 재표본 방법을 이용한 GARCH(1,1) 모형의 예측, 동국대학교.

[17] Brocklebank, Dickey(2003). SAS for Forecasting Time Series, SAS.

[18] William W.S. Wei(2006), TIME SERIES ANALYSIS Unicariate and Multivariate Methods Second Edition, Pearson International Edition.

[투고일: 2009. 08. 10][심사(수정)일: 2009. 08. 15][게제확정일: 2009. 08. 19]

Annals of Statistical Research
Vol. 6. No. 1, 2009, pp.57~69

An Algorithm for Calculating Normal Screening Intervals

In-Ki Kim[1], Hea-Jung Kim[2]

Abstract

This article develops a class of the weighted normal distributions for which the probability density function has the form of a normal density and a weighted function. The form of this class is obtained from the conditional distribution of a doubly truncated normal distribution. Therefore, this distribution is useful for the screening.

HPD interval is of the shortest length among all possible credible intervals. Using Chen-Shao algorithm, we calculate a truncation points of HPD interval.

Keywords: Weighted normal distribution; Truncated bivariate normal; Screening; Constrained analysis; HPD; Chen-Shao

1) Doctoral Candidate for Statistics, Department of Statistics, Dongguk University, Seoul 100-715, Korea. E-mail : inkey@dongguk.edu
2) Corresponding Author, Professor, Department of Statistics, Dongguk University, Seoul 100-715, Korea. E-mail : kim3hj@dongguk.edu

1. 서 론

최근에 들어 절단 이변량 정규분포(truncated bivariate normal distribution)의 주변분포는 응용분야의 증가에 따라 이에 대한 연구도 활발히 진행되어 왔다. Azzalini(1985)에 의해 비대칭정규분포 (skew-normal distribution)가 제안한 이래 Bayarri and DeGroot(1992), Branco and Dey(2001), Amold and Beaver(2002), Ma et al.(2005), 김혜중(2005)등은 비대칭 정규성을 가진 다양한 분포들을 제안하였다. 이 분포들은 가중분포족 (weighted distribution family)에 포함되는 것들이며, 가중분포는 잠재 변수 X의 확률밀도함수 $g(x;\theta_1)$에 어떤 모수 벡터 θ_1와 θ_2를 갖은 비음가중 함수 $w(x;\theta_1,\theta_2)$를 곱하여 유도되는 분포이다.

이 가중분포의 확률밀도함수는 다음과 같이 형태를 갖는다.

$$f(x;\theta_1,\theta_2)=\frac{w(x;\theta_1,\theta_2)g(x;\theta_1)}{E_{\Theta_1}[w(X;\theta_1,\theta_2)]}. \qquad (1)$$

특히 식(1)에서 정의된 $g(x;\theta_1)$가 정규밀도함수일 경우 $f(x;\theta_1,\theta_2)$를 가중정규분포(weighted normal distribution)라 한다. 그리고 가중정규분포의 한 형태는 이변량이중절단정규분포의 조건부 분포로부터 얻어질 수 있다. 김혜중(2007)은 이변량이중절단정규분포를 사용하여 새로운 형태의 가중정규분포를 제안하였으며, 가중정규분포는 Boys et al.(1986)이 고려한 선별(screening) 표본을 계획하는데 적절한 모형이 될 수 있다. 선별 표본법이란 $P(Y\in C_Y|X\in C_X)=\alpha$일 경우 C_X가 주어졌을 때 C_Y가 최단 길이 인 특정한 영역 내에서 표집하는 표본법이다. 여기서 C_X와 C_Y는 각각 확률변수 X, Y의 확률표집구간을 나타낸다. 확률변수 Y의 값이 특정한 영역C_Y 에 속하면 개체 측정이 성공이라고 하자. . 그러면 $\gamma=P(Y\in C_Y)$은 성공할 확률 또는 모집단 속의 성공할 개체수이다. 선별 표본법이란 확률변수 Y의 측정이 측정비용 및 시간등의 이유로 어려울 경우 확률변수 Y와 상관관계가 높고 측정 비용이 저렴하고 측정시간이 짧은 확률변수X의 성공여부를 먼저 조사한 후 확률변수 Y를 조사함으로서 측정의 성공확률을 높이는데 목적이 있다.

C_X와 C_Y가 단측 영역을 가졌을 때의 선별방법에 대하여는 Owen et al(1981), Boys and Dunsmore(1986)등이 연구하였고 더 일반적인 C_X와 C_Y가 양측영역을 가질 경우는 Li and Owen(1979), Riew(1985)등이 연구하였으나 C_X가 $(a\le X\le b)$일 경우만을 언급하였다.

본 논문은 가중정규분포를 사용하여($a \le X \le b$)일 경우 뿐 만 아니라 ($X < c,\ d < X$)일 경우의 C_Y를 구하는 방법에 대하여 연구하였다. 이를 위해 양측절단이변량분포와 중심절단이변량분포로부터 유도된 두 종류의 가중정규분포를 제시하고 이들 각각의 속성 및 확률적 표현에 대하여 연구하였다. 또한 유도된 분포로 Chen-Shao 알고리즘을 사용하여 수치적으로 C_Y를 구하는 방법에 대해서 설명하였다.

2. 가중정규분포군

식(1)에 의하여 $X_1 \sim N(\mu, \sigma^2)$ 일 경우 X_1 의 pdf가 비음가중함수 $w(x:\theta)$ 를 곱하여져 기울어졌다면 가중정규분포군은 다음과 같이 정의된다.

$$f_{X_1}(x) = \sigma^{-1}\phi\left(\frac{x-\mu}{\sigma}\right)\frac{w(x:\theta)}{E[W(X_1:\theta)]} \qquad x \in R \tag{2}$$

여기서 ϕ 는 표준정규밀도함수(pdf)이고 $\theta(\{\mu,\sigma\} \subset \theta)$는 가중함수에 포함된 모수 백터를 표시한다.

식(2)의 형태를 갖는 가중정규분포는 양측절단정규분포의 조건부 분포로부터 얻어진다. (X_1, Y)을 평균 벡터가 (μ_1, μ_2), 분산 벡터가(σ_1^2, σ_2^2)이고 상관계수가 ρ인 정규이변량이라 하자. 확률변수 $X = [X_1 | a < Y < b]$라고 정의하면 적분에 의하여 X의 밀도함수는 다음과 같이 주어진다.

$$f_X(x) = \sigma^{-1}\phi(z)\frac{\Phi(\delta u(b) - \lambda z) - \Phi(\delta u(a) - \lambda z)}{\Phi(u(b) - \Phi(u(a))}, \qquad x \in R \tag{3}$$

여기서 a, b는 실상수이고 $z = (x-\mu_1)/\sigma_1,\ \mu(a) = (a-\mu_2)/\sigma_2,\ \mu(b) = (b-\mu_2)/\sigma_2,$ $\delta = 1/\sqrt{1-\rho^2},\ \lambda = \rho/\sqrt{1-\rho^2}$이다.

$Z \sim N(0,1)$일 때, 모든 실수 c, d 에 대하여 $E[\Phi(cZ+d)] = \Phi\left(\frac{d}{\sqrt{1+c^2}}\right)$이고, 모든 실수 a에 대하여 $E[\Phi(\delta a - \lambda Z)] = \Phi(a)$이므로 양측절단이변량정규분포의 주변분포(3)

은 가중함수 $w(x;\theta)=\Phi(\delta u(b)-\lambda z)-\Phi(\delta(u(a)-\lambda z),\ \theta=(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)'$인 가중정규분포(2)로 된다. 만약 X의 pdf가 식(3)과 같다면. 확률변수 X는 모수 벡터 $\theta=(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)'$와 절단점$(a,b)$를 가진 가중정규분포를 갖고 있다고 말할 수 있다.이것을 $X\sim\ WN_{(a,b)}$ 라고 표기하도록 한다.

$WN_{(a,b)}$분포는 독립적인 정규분포와 양측절단정규분포의 합으로 표시할 수 있다. $U^{(1)}\sim N(\theta_1,\tau_1^2)$ 와 $U^{(2)}\sim N(\theta_2,\tau_2^2)$ 이 독립변수라 하면 $\rho=c_2\sigma_2/\sigma_1$ 을 만족하는 실수 $c_1(\neq 0)$과 c_2에 대하여 다음 관계가 성립한다.

$$c_1U^{(1)}+c_2U^{(2)}_{(a,b)}\sim\ WN_{(a,b)}(\theta) \tag{4}$$

여기서 $\theta=(\mu_1,\mu_2,\sigma_1,\sigma_2)',\ \mu_2=\theta_2,\ \sigma_1^2=\sum_{i=1}^{2}c_i^2\tau_1^2$이고 $\sigma_2^2=\tau_2^2$이다.

따라서 $Z_1\sim N(0,1)$와 $Z_2\sim N(0,1)$가 독립이라 하면 식(4)에 의해

$$\frac{1}{\sqrt{1+\lambda^2}}Z_1+\frac{\lambda}{\sqrt{1+\lambda^2}}Z_{2(a,b)}\sim\ WN_{(a,b)}(\theta) \tag{5}$$

여기서 $\lambda=\rho/\sqrt{1-\rho^2}$이다.

확률변수 Y가 절단점$(a,\ b)$에서 양측으로 절단될 수도 있지만 $(Y<a,\ b<Y)$와 같이 종심절단이 될 경우도 있다.

$X=[X_1|\,Y<a,\ b<Y]$ 이라 하면 $f_X(x)=\dfrac{\int_y f_{X,Y}(x,y)dy}{f_Y(y)}$이므로 직접 적분을 하면,

분자 $\int_y f_{X,Y}(x,y)dy=\int_{-\infty}^{a}f_{X,Y}(x,y)dy+\int_b^{\infty}f_{X,Y}(x,y)dy$

$$=\sigma^{-1}\phi(z)\int_{-\infty}^{\delta u(a)-\lambda z}\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}w^2)dw+\sigma^{-1}\phi(z)\int_{\delta u(b)-\lambda z}^{\infty}\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}w^2)dw$$

$$= \sigma^{-1}\phi(z)\left(1 - \int_{\delta u(a)-\lambda z}^{\delta u(b)-\lambda z} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}w^2)dw\right)$$

$= \sigma^{-1}\phi(z)[1 - ((\Phi(\delta u(b) - \lambda z) - (\Phi(\delta u(b) - \lambda z))]$이 되고,

분모 $f_Y(y) = P[Y < a,\, b < Y] = 1 - P[a < Y < b]$ 이므로

$$f_X(x) = \sigma^{-1}\phi(z)\frac{1 - [\Phi(\delta u(b) - \lambda z) - \Phi(\delta u(a) - \lambda z)]}{1 - [\Phi(u(b)) - \phi(u(a))]}, \qquad x \in R \tag{6}$$

이다.

$X = [X_1 | Y < a,\ b < Y]$의 pdf는 식(6)과 같으며 이는 가중정규분포 (2)와 비교해 볼 때

$$w(x;\theta) = 1 - [\Phi(\delta u(b) - \lambda z) - \Phi(\delta(u(a) - \lambda z)]$$

$$E(w(x;\theta)) = 1 - [\Phi(u(b) - \Phi(u(a))]$$

임을 쉽게 알 수 있다. 따라서 X의 분포는 가중정규분포이다.

X의 pdf가 식(6)과 같다면. 확률변수 X는 모수 벡터 $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)'$와 절단점 (a, b)를 가진 가중정규분포를 갖고 있다고 말할 수 있다. 이것을 $X \sim WN_{(\backslash ab)}$ 라고 표기하도록 한다. $Z_1 \sim N(0,\ 1)$ 와 $Z_2 = N(0,1)$가 독립일 경우 식(5)과 근사하게

$$\frac{1}{\sqrt{1+\lambda^2}} Z_1 + \frac{\lambda}{\sqrt{1+\lambda^2}} Z_{2(\backslash ab)} \sim WN_{(\backslash ab)}(\theta) \tag{7}$$

이다.

이렇듯 $WN_{(a,b)}$, $WN_{(\backslash ab)}$모두 두 분포의 합으로서, 또 ρ의 함수로서 표시할 수가 있다.

3. 가중정규분포군의 응용

선별표본에 대한 문제는 여러 논문에서 중요하게 다루어졌다 (Ma et al.(2005)와 거기에 수록된 참고문헌을 참조). 어떤 개체의 측정치가 정규 확률변수 Y에 의하여 행하여진다고

하자. 어떤 특정한 영역 C_Y에 $Y \in C_Y$이면 개체측정이 성공이라고 하자. 그러면 $\gamma = P(Y \in C_Y)$은 성공할 확률 또는 모집단 속의 성공할 개체수이다. C_Y의 상하한 영역을 $[L, U]$로 표기하기로 하자.

선별하려는 이유는 몇 개의 개체를 소거시켜 성공 비율을 증가시키기 위함이다. 선별은 각 개체의 두 번째 측정치 X 를 조사함으로서 이루어진다. X와 Y는 상관관계가 있으며 정규확률변수 X는 Y보다 먼저 측정되어 진다. X가 Y보다 측정하기가 용이하거나, 측정비용이 저렴하거나, 선행되어지는 절차일 경우 등이다.

선별 표본에서 빈번히 발생하는 문제는 C_X가 주어졌을 때 특정한 영역 C_Y의 최단 길이를 결정하는 것이다. C_X는 X의 특정한 영역으로 $X \in C_X$ 이던지 아닌가에 따라 각각 남아있던지 선별되어 진다. 목표는 C_Y의 최단 길이를 찾는 것으로$P[L < Y < \mathrm{U} | a < X < b)] = \alpha$ 또는$P[L < Y < \mathrm{U} | X < a,\ b < X)] = \alpha$라 하면 $\alpha > \gamma$이다.

중요한 변수가 있고, 이 변수는 측정될 수 있으며, 선별되지 않은 원 자료로부터 $(y_1, x_1),\ (y_2, x_2), \cdots, (y_n, x_n)$ 의 형태인 확률변수를 추출하여 이용할 수 있다고 하자. $C_Y = [L, U]$ 의 최단 영역은 $P[L \le Y \le U | a \le X \le b] = \alpha$ 또는 $P[L \le Y \le U | a \le X \le b] = \alpha$경우와 같이 양측 영역이 요구된다.

제약조건하에서 $U - L$ 의 길이를 최단화하는 L과 U를 발견하는 것이 관건이다.

$$F_Y(U) - F_Y(L) = \int_L^U f_Y(y)dy = \alpha \text{ 이고}$$

여기서 $F_Y(y)$와 $f_Y(y)$ 는 각각 $WN_{(a,b)}(\theta)$와 $WN_{(\backslash a,b)}$분포의 df와 pdf이다.

U 와 L 사이가 최단 영역이 되려면

$F_Y(U) - F_Y(L)$을 U로 미분하면 $\dfrac{d(F_Y(U) - F_Y(L))}{dU} = f_Y(U) - \dfrac{F_Y(L)}{dL}\dfrac{dL}{dU} = 0,$ 따라서

$\dfrac{dL}{dU} = \dfrac{f_Y(U)}{f_Y(L)}$이고, U 와 L 사이가 최단 영역이 되려면 $\dfrac{d(U-L)}{dU} = 0$을 만족시켜야하므로

$\dfrac{dL}{dU} = 1$으로 $1 = \dfrac{f_Y(U)}{f_Y(L)}$으로 U 와 L은 $f_Y(U) = f_Y(L)$ 인 조건을 만족시켜야한다.

즉 $f_Y(U) = f_Y(L)$인 조건하에서 $P(L \le Y \le U | a \le X \le b) = \alpha$ 또는

$P[L < Y < \mathrm{U} | X < a,\ b < X)] = \alpha$ 를 찾아야 한다.

그러나 (3), (5), (6), (7)과 같은 분포를 알고 있다고 하여도 최단 확률구간 (L, U)를 구하는 것은 어려운 일이다.

4. 최단확률구간의 수치적 계산 알고리즘

$\pi(\theta|D)$와 $\Pi(\theta|D)$를 각각 주변사후밀도함수와 주변사후누적함수(cdf)라 하면 θ에 대한 100(1-α)% 신용구간(credible interval)은 다음과 같은 형태로 표현되고 Θ의 사후분포를 사용했기 때문에 확률구간이다.

$$(\theta^{(\alpha/2)}, \theta^{(1-\alpha/2)}),$$

여기서 $\Pi(\theta^{(\alpha/2)}|D = \alpha/2, \Pi(\theta^{(1-\alpha/2)}|D) = 1-\alpha/2$이다.

$100(1-\alpha)\%$ 포함확률을 가지는 θ의 신용구간 중 크기가 가장 작은 신용구간을 HPD(highest posterior density)신용구간이라 한다. 이것은 다음과 같이 표현된다.

$$R(\pi_\alpha) = \{\theta : \pi(\theta|D) \geq \pi_\alpha\} \qquad (8)$$

π_α는 $P(\theta \in R(\pi_\alpha)) \geq 1-\alpha$를 만족하는 최대의 상수이다.

전형적으로 모수 θ 에 대한 사후분포가 단봉일 때 HPD는 하나의 구간이지만 사후분포가 이봉분포이면 HPD 신용구간은 서로 겹치지 않는 두 구간의 합이 된다.

HPD 구간을 계산하는 것은 신용구간을 계산하는 것보다 더욱 더 어렵다. 식(9)에서 정의된 $R(\pi_\alpha)$를 구하는 것은 $\pi(\theta|D)$와 $\Pi(\theta|D)$의 근접한형태(closed form)를 알고 있다 하더라도 간혹 유효하지 못하다.$R(\pi_\alpha)$을 계산하기 위해서는 π_α를 알고 식(8)에 의하여 구해야하지만 π_α를 알기 어렵기 때문이다. 근사적인 방법으로 π_α를 $P(\xi \geq \pi_\alpha) = 1-\alpha$ 일 때의 ξ의 α^{th} 백분위수로 구하는 방법이 Wei 와 Tanner[13], Hyndman[14]에 의하여 개발되어 왔다. 약점은 $\pi(\theta|D)$의 근접한형태를 알아야 한다는 것이고 $\widehat{\pi_\alpha}$를 추정한다 할 지라도 $\widehat{R(\pi_\alpha)}$를 계산하기 어렵다는 것이다.

이러한 단점을 보안하기 위하여 Chen-Shao[15]가 제안한 방법은 근접한형태에 불문하고 계산할 수 있는 장점과 $\widehat{R(\pi_\alpha)}$계산이 용이하다는 장점이 있으며 Chen-Shao 알고리즘을

이용하여 HPD를 구하는 절차는 다음과 같다.

Step 1. 식(5)의 $WN_{(a,b)}(\theta)$, 또는 식(6)의 $WN_{(\backslash ab)}(\theta)$ 분포로부터 generating에 의한 표본$\{x_i, i=1,2,\cdots,n\}$를 얻는다.

Step 2. $\{x_i, i=1,2,\cdots,n\}$를 정렬시켜 순서화 된 값 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$을 얻는다.

Step 3. $100(1-\alpha)\%$ 신용구간을 계산한다.

$$R_j(n) = \{x_{(j)},\ x_{(j+[(1-\alpha)n])}\}$$

$$\text{for } j = 1,\ 2,\ \cdots,\ n-[(1-\alpha)n]$$

Step 4. 모든 신용구간 중에 가장 작은 길이를 가진 HPD를 정한다.

$WN_{(a,b)}$ 가 이봉형이면 최대 두 개 구간의 합으로 HPD가 계산될 것이다. 첫 번재 구간은 첫 번째 봉우리 구역 내 존재하고 두 번째 구간은 두 번째 봉우리 구역 내 존재하게 된다.

첫 번째 구간의 시작점을 i, 간격이 m 따라서 끝점은 $i+m$이고, 두 번째 구간의 시작점을 j라고 하자. m이 가질 수 있는 최대 값은 분포의 바른편 끝 쪽에 αn개가 있을 때이다. j는 두 번째 구간의 시작점이므로 첫 번째 구간의 끝점인 $i+m$와 같거나 커야한다. 또한 두 번째 구간의 최대 간격은 αn을 제외한 이외에도 첫 번째 구간의 간격 m을 제하여야 한다. 따라서 j의 최대 값은 $j+\alpha n+m=n$을 만족시켜야 한다. i의 범위는 j의 범위에서 자동적으로 계산된다. 따라서 HPD는 아래와 같이 계산된다.

$$HPD = \min_{0 \le m \le (1-\alpha)n} \min_{0 \le i \le n-[(1-\alpha)n]-2m} \min_{i+m \le j \le n-[1-\alpha)n]-m} \times \{(x_{(i+m)} - x_{(i)}) + (x_{(i+[(1-\alpha)n]-m)} - x_{(j)})\}$$

HPD의 영역은 $(x_{(i+m)},\ x_{(i)}) \cup (x_{(j+[(1-\alpha)]n-m)},\ x_{(j)})$ 이 된다.

또한 HPD의 영역에 속하는 표본의 수는 $(i+m)-i+j+[(1-\alpha)]n-m-j=(1-\alpha)n$이다.

5. 알고리즘 적용 결과와 해석

표본 추출 확률 α가 정해지고 선행으로 조건 지워지는 변수 X의 영역과 절단점이 정해지면 즉 $(a<X<b)$ 혹은 $(X<a,\ b<X)$에서 $(a,\ b)$가 정해지면 목표변수 Y의 최단거리를 갖는 절단점 L과 U를 찾는 것이 목표이다.

$P[L \le Y \le U \mid a \le X \le b] = \alpha$ 일 때 최단의 $U-L$을 Chen-Shao의 알고리즘으로 구한 결과는 <표 1>이고 $P[L \le Y \le U \mid X < a, b < X] = \alpha$ 일 때의 최단의 $U-L$을 Chen-Shao의

알고리즘으로 구한 결과는 <표 2>이다.

Chen-Shao 알고리즘을 적용시 2,000개의 표본을 generating 하였다.

선별 표본했을 때의 확률은 α(=$P[L \leq Y \leq U | a \leq X \leq b]$또는

$P[L \leq Y \leq U \mid X < a, b < X]$)

이고 선별 표본하지 않았을 때의 확률은 $P[L \leq Y \leq U] = \gamma$이다.

선별 표본의 효과는 α와 γ를 비교해보면 용이하게 알 수 있다.

<표 1> $P[L \leq Y \leq U | a \leq X \leq b] = \alpha$**일 경우 주어진**$(a,\ b)$, α, ρ**에 대해 계산된** $(L,\ U)$,γ**값**

a	b	α	ρ	L	U	$U-L$	γ
-1	1	0.9	0.10	-1.6465	1.5899	3.2364	0.8942
			0.30	-1.5529	1.5821	3.1350	0.8830
			0.50	-1.5125	1.4186	2.9311	0.8568
			0.80	-1.3497	1.0638	2.4135	0.7677
			0.90	-1.0854	1.0191	2.1045	0.7070
			0.95	-0.9586	0.9706	1.9292	0.6652
-1.5	1	0.8	0.10	-1.2705	1.1525	2.4230	0.7735
			0.30	-1.2152	1.1287	2.3439	0.7584
			0.50	-1.1339	1.0904	2.2243	0.7338
			0.80	-1.0327	0.8907	1.9234	0.6626
			0.90	-0.9988	0.8049	1.8037	0.6306
			0.95	-1.0004	0.7660	1.7664	0.6196
-2	1	0.8	0.10	-1.3616	1.1925	2.5541	0.7968
			0.30	-1.2371	1.2339	2.4710	0.7834
			0.50	-1.1835	1.1523	2.3358	0.7571
			0.80	-1.2821	0.8220	2.1041	0.6945
			0.90	-1.0857	0.8906	1.9763	0.6746
			0.95	-1.0894	0.8390	1.9284	0.6613

<표 2> $P[L \le Y \le U | X < a, b < X] = \alpha$ **일 경우 주어진** $(a,\ b)$, α, ρ**에 대해 계산된** $(L,\ U)$,γ**값**

a	b	α	ρ	R				D	γ	model
				L_a	L_b	U_c	U_d	$L_b - L_a + U_d - U_c$		
-1	1	0.9	0.10	-1.7548	-	-	1.5490	3.3038	0.8997	단봉
			0.30	-1.7627	-	-	1.7237	3.4863	0.9186	//
			0.50	-1.9350	-	-	1.8183	3.7533	0.9390	//
			0.80	-2.1173	-0.0323	0.4942	2.0451	4.0807	0.9299	이봉
			0.90	-2.3111	-0.2465	0.4605	2.1836	3.7877	0.7003	//
			0.95	-2.1557	-0.5432	0.5337	2.2473	3.3261	0.5624	//
-1.5	1	0.8	0.10	-1.0820	-	-	1.4418	2.5238	0.7857	단봉
			0.30	-1.0102	-	-	1.6773	2.6875	0.7971	//
			0.50	-1.3772	-	-	1.7195	3.0967	0.8730	//
			0.80	-2.0291	-0.9393	0.1646	2.3495	3.2746	0.5778	이봉
			0.90	-2.0924	-1.2762	0.4223	2.3411	2.7350	0.4095	//
			0.95	-2.2104	-1.4211	0.6485	2.2167	2.3575	0.3091	//
-2	1	0.8	0.10	-1.3832	-	-	1.2676	2.6508	0.8142	단봉
			0.30	-0.9324	-	-	1.7995	2.7319	0.7885	//
			0.50	-0.9084	-	-	1.9252	2.8336	0.7911	//
			0.80	-0.1102	2.3144	2.4456	2.5270	2.5061	0.5350	이봉
			0.90	0.2882	0.3545	0.4010	2.4098	2.0751	0.3613	//
			0.95	0.5314	2.2739	2.3764	2.4546	1.8208	0.2878	//

L_a와 L_b는 각각 아래봉우리의 아래 절단값과 위 절단값이고 U_c과 U_d는 각각 위 봉우리의 아래 절단값과 위 절단값을 나타낸다. D는 최단확률구간의 길이를 나타낸다.

출력 결과를 요약하면 다음과 같다.

(1) α와 γ를 비교해 보면 ρ가 커질수록 γ 가 작아져 제약조건식의 HPD가 작아진다.

1) $P(L \le Y \le U | a \le X \le b)$인 제약식을 가진 경우 ρ가 커질수록 γ 가 급격히 작아지지는 않는다.

절단의 비대칭율이 커질수록 γ 가 작아지지는 않는다.

모든 경우 γ 가 α보다 작다.

2) $P[L \le Y \le U | X < a, b < X] = \alpha$일 경우 ρ가 커질수록 γ 가 급격히 작아진다.

절단점 a, b가 비대칭 일수록 γ 가 작아지며 분포 형태가 이봉이 될 때 급격히 작아진다.

모든 경우에 γ 가 α보다 작지는 않다. 오히려 절단점 a, b의 비대칭률이 적을 경우 γ

가 α보다 크며 상관계수 ρ가 어느 정도 커질 때 (0.5, 0.6 부근)까지는 γ 가 점점 커진다.

(2) $P[L \leq Y \leq U | a \leq X \leq b] = \alpha$인 제약식을 가진 모든 경우 γ 가 α보다 작지마는 $P[L \leq Y \leq U | X < a, b < X] = \alpha$ 인 제약식인 경우 모든 경우에 γ 가 α보다 작지는 않다. 오히려 절단점 a, b의 비대칭률이 적을 경우 γ 가 α보다 크며 상관계수 ρ가 어느 정도 커질 때 (0.5, 0.6 부근)까지는 γ 가 점점 커진다.

(3) 비대칭률이 커지고 ρ 가 커질수록 γ 가 작아진다. 특히 분포형태가 이봉이 될 때부터 급격히 γ 가 작아진다.

5. 결 론

본 논문는 가중정규분포의 일족인 양측절단정규분포의 조건부 분포와 중심절단정규분포의 조건부 분포로 구성되어진 이중절단정규분포의 조건부 분포의 속성과 응용에 대하여 연구한 것이다. 이중절단정규분포의 조건부 분포를 독립적인 정규분포와 이중절단정규분포의 합으로 표기하여 성질을 알아보았고 주변확률분포도 알아 보았다. 몇 개의 개체를 소거시켜 성공 비율을 증가시키기 위한 선별 표본의 문제는 여러 영역에서 응용되질 수 있다. 특히 선별 표본에서 빈번히 발생하는 문제는 선행으로 조사되어지는 변수의 영역 C_X가 주어졌을 때 특정한 영역 C_Y의 최단 길이를 결정하는 것이다. 이 분포는 자체 속성상 선별 문제를 해결하는데 적합한 분포이다. 그러나 분포의 형태를 알고 문제의 해결 조건을 안다고 하여도 그 과정은 어려운 것이다. 그래서 Chen-Shao 알고리즘을 이용하여 generating에 의하여 선별 표본을 실시하여 추출확률을 결정하고 선행으로 조건 지워지는 변수의 절단점과 영역이 확정되었을 때의 최단확률구간 (L, U)을 구하였다. 선별 표본한 결과를 보면 조건 지워지는 변수의 절단점의 비대칭률이 커지고 두 변수가 상관계수가 클수록 선별 표본 효과가 크고, 특히 분포형태가 이봉이 될 때부터 급격히 선별 표본 효과가 증대해지는 것을 알 수 있다. 즉 조건 지워지는 변수의 영역 C_X가 중심절단 되었을 때 아주 효과적이다.

참고문헌

[1] Amold, B.C. and Beaver, R.J., (2002). Skewed multivariate models related to hidden truncation and/or selective reporting. Test, 11,7-54.

[2] Azzalini, A., (1985). A class of distributions which includes the normal ones. Scandinavian Journal of Statistics, 12. 171-178.

[3] Bayarri, M.J. and DeGroot, M., (1992). A BAD view of weighted distributions and selection models, Bayesian Statistics, Vol.4(Oxford:Oxford University Press).

[4] Boys, R.J. and Dunsnore, I.R., (1986). Screening in a normal model. Joural of Royal Statistical Society Series B, 48, 60-69.

[5] Branco, B.C. and Day, D.K., (2001). A general class of multivariate skew-normal distribution. Journal of Multivariate Analysis, 79, 99-113.

[6] Chen, M.H., Shao, Q.M., Ibrahim. J.G., (2001). Monte Carlo Methods in Bayesian Computation(New York:Springer)

[7] Genton, M.G, (2005). Discussion of "the skew-normal". Scandinavian Journal of Statistics,32, 189-198.

[8] Hyndman, R.J., (1996). Computing and Graphing Highest Density Regions, The American Statistician, 50, 120-126.

[6] Kim, H.J., (2005). On a class of two-piece skew-normal distributions. Statistics, 39,537-553.

[9] Kim, H.J., (2007), A class of weighted normal distributions and its variants useful for inequality constrained analysis, Statistics, 41, 421-441.

[10] Li, L., Owen, D.B., (1979). Two-Sided Screening Procedures in the Bivariate Case, Technometrics, 21, Truncated Normal, The Korean Communications in Statistics, 13, 255-266.

[11] Ma, Y., Genton, M.G. and Tsiatis, A.A., (2005). Locally efficient semiparametric estimators for generalized skew-elliptical distributions. Journal of the American Statistical Association 100, 980-989.

[12] Owen, D.B., Li, l. and Chou, Y.M., (1981) Prediction Intervals for Screening Using a Measureed Correlated Variate, Technometrics, 23, 165-199.

[13] Riew, M.C., (1985) Optical Screening Procedures for Improving Outgoing Quality Based on Correlateded Normal Variables, Journal of the Korean Statistical Society, 14,18-28

[14] Wei, G.G. and Tanner, M.A., (1990). Culculating the Content and Boundary of the Highest Posterrior Density Region via Data Augmentation, Biometrika, 77, 649-652.

[투고일: 2009. 08. 13][심사(수정)일: 2009. 08. 18][게제확정일: 2009. 08. 19]