응용통계연구 (2009)
22(5), 981–993

대졸자 직업이동 경로조사에서 패널탈락분석

천영민[1] · 윤정혜[2] · 오민홍[3]

[1]한국고용정보원 고용조사분석센터, [2]한국고용정보원 고용조사분석센터,
[3]한국고용정보원 인력수급전망센터

(2009년 7월 접수, 2009년 8월 채택)

요 약

패널조사에서 패널탈락이 특정계층에 집중되어 있다면, 초기에 구축된 샘플의 변화에 따라 패널자료의 대표성에 문제를 야기할 수 있다. 본 연구는 대졸자 직업이동 경로조사(GOMS)를 이용하여 신뢰성과 대표성을 저해하는 표본탈락편의(non-random attrition bias)가 있는지를 파악하고, 패널탈락의 결정요인을 분석하여 패널탈락을 최소할 수 있는 방안을 모색하고자 한다. 분석결과 패널탈락은 응답자의 문제보다 조사시스템의 문제가 더 큰 것으로 나타났다. 따라서 추가연구를 통해 체계적인 조사시스템의 구축 및 응답자 관리방법 개발뿐만 아니라 패널탈락의 편의를 보완하기 위한 가중치 부여 등 다양한 개선책의 도입이 시급한 것으로 판단된다.

주요용어: 로짓 모형, 의사결정나무, 패널 탈락, CAPI.

1. 서론

패널조사(panel survey)는 특정 시점의 전체적인 상황에 대한 분석을 할 수 있는 횡단면 조사(cross-sectional survey)와 시간의 흐름에 따라 변해가는 현상을 추적하는 종단면 조사(longitudinal survey)를 모두 활용할 수 있다는 측면에서 각광을 받고 있다. 국내에 대우패널이 처음 도입된 이후, 수많은 패널조사들이 진행되고 있으며, 현재에도 여러 기관에서 다양한 패널조사를 기획 및 준비하고 있다. 반면에 어떤 패널조사들은 이미 중단되어 버리기도 했는데, 그 이유는 응답대상이 소진되어 패널조사의 목적이 상실되었거나, 해를 거듭하며 조사된 결과가 정책적으로 활용되기에 미흡하다는 판단 때문이다. 무응답(non-response)은 조사를 설계하고 기획한 이들이 응답자로부터 원하는 자료를 얻지 못하는 것을 말한다. 패널조사에서 무응답에 대한 연구가 진행되기 이전의 무응답 발생 형태는 단위 무응답(unit non-response)과 항목 무응답(item non-response)으로 나뉘었다. 단위 무응답은 표본(sample)으로 선택된 대상(object) 또는 개체(case)로부터 조사를 실시하지 못해 발생하는 것으로, 이에 대한 처리는 대체 표본을 통해 해결하게 된다. 이에 반해 항목 무응답은 조사 대상으로부터 어떤 특정한 문항에 대한 정보를 얻지 못해서 발생하는 것이다. 이는 대부분의 조사에서 발생하는 것이며, 이를 얻기 위해 강제하는 것은 다른 문항에까지 영향을 미치기 때문에 무응답 형태로 조사를 완료한 후에 대체(imputation)를 통해 보정하거나 무응답이 존재하는 자료 형태로 분석하게 된다. 패널조사에 대한 무응답 대체 연구가 시작된 이후에 패널 무응답(panel non-response) 또는 웨이브 무응답(wave non-response)에 관심을 갖기 시작했다. 패널 무응답이란, 조사 대상인 패널이 특정년도에 응답을 하지 않다가 다시 패널조사에 응답을 하는 형태로 특정년도에만 마치 단위 무응답의 형태를 보이기 때문에,

[3]교신저자: (150-836) 서울시 영등포구 문래동 3가 77-11, 한국고용정보원 인력수급전망센터, 연구위원.
E-mail: mhoh7266@keis.or.kr

횡단면 연구에서는 큰 문제가 되지 않지만 종단면 연구에서는 대상의 포함여부가 문제가 된다. 물론 불균형 패널(unbalanced panel) 형태로 자료를 분석하는 방법이 있지만 여전히 분석의 한계점이 존재하고, 자료의 완전성(completeness) 또는 대표성에 대한 문제제기가 될 뿐만 아니라, 이런 패널의 탈락이 특정한 계층이나 사회적인 현상으로 발생하는 것이라면, 패널의 장기적 관리 측면에서 큰 손실이 아닐 수 없다. 패널조사 및 조사된 결과의 신뢰성과 대표성은 표본, 조사방식, 표본유지율 등을 통해 확인하게 된다. 표본은 모집단을 얼마나 가장 잘 대표할 수 있게 선정되었는지 그리고 표본추출방식은 신뢰할 만한 것인지에 대한 것이다. 조사방식은 일반적인 1회성 횡단면 조사와는 다른 방식을 통해 조사를 해야 하는 것을 의미한다. 전년도에 조사된 일부 항목의 경우, 변동 또는 추적의 필요가 없는 경우라면, 그 항목에 대해 다시 질문을 하면 안 될 것이며, 전년도에 조사된 기본정보를 바탕으로 접근할 수 있느냐 하는 것이다. 따라서 PAPI(paper and pencil interview) 조사 방식 보다는 CAPI(computer assisted personal interview) 방식을 최근 패널조사에서 선호하고 있다. 표본유지율은 각 패널조사에서 가장 많이 신경을 쓰는 것으로, 패널조사 결과의 지수(index) 역할을 한다. 표본유지율이 어느 정도인가에 따라 그 조사가 얼마나 성실하게 진행되었는지 또는 조사결과가 얼마나 신뢰할 만한 것인지를 판단하게 된다. 표본유지율 제고를 위해 탈락되는 패널들에 대한 특성을 파악하는 것은 중요한 문제가 되었다. 탈락되는 패널들이 어떤 특정한 요인들에 의해 발생되는 것이 탐지된다면, 이 문제를 해결하여 표본유지율을 향상시킬 수 있을 것이기 때문이다. 뿐만 아니라 탐지된 요인들 중에 해결이 불가능한 문제가 있다면, 다른 패널조사들에게 일종의 신호기제의 효과를 줄 수도 있을 것이다.

대졸자 직업이동 경로조사(graduates occupational mobility survey, 이하 GOMS)에서 패널탈락(panel attrition)이 특정 계층이나 경제활동상태에 집중되어 있다면, GOMS가 대졸 청년층의 노동시장 특성을 대표할 수 있는 자료로서 가치에 손상을 입힐 수 있다. 따라서 GOMS에서 신뢰성과 대표성을 저해하는 패널탈락 편의(non-random attrition bias)가 존재한다면, 패널탈락의 결정요인과 패널탈락이 주요 변수에 어떤 영향을 미치는지 파악하고 이를 검증하고자 한다. 뿐만 아니라 향후 조사에서 이를 기반으로 패널탈락을 최소화 할 뿐만 아니라 탈락된 패널을 다시 복귀시키기 위해서 어떤 방안을 강구해야 하는지 등에 대해서 알아보고자 한다.

본 연구의 구성은 다음과 같다. 1장에서는 서론으로 연구의 목적 등에 대해 설명하고, 2장에서는 패널탈락과 관련한 선행 연구에 대해 살펴보고, 3장에서는 본 연구에서 사용한 대졸자 직업이동 경로조사 자료에 대해 소개하고, 4장에서는 분석에 사용된 변수와 분석 방법 등에 대해 설명하고, 5장에서는 분석결과, 6장에서는 결론 및 본 연구의 성과에 대해 설명한다.

2. 패널탈락에 대한 선행 연구

Becketti 등 (1998)은 1968년부터 1981년까지의 14년간 PSID(panel study of income dynamics) 자료를 이용하여 표본의 이탈률을 점검하였다. 분석 결과, PSID의 자체표본인 SRC(survey research center) 표본보다 빈곤가구로 추출된 표본인 SEO(survey of economic opportunity) 표본의 이탈률이 높은 것으로 나타났다. Lillard와 Panis (1998)는 1968년부터 1988년까지 21년간의 PSID 자료를 이용하여 응답확률을 분석하였는데, 백인 남성과 혼인한 지 오래된 기혼부부가 응답확률이 높은 것을 보여주었다. Fitzgerald 등 (1998)은 1968년부터 1986년까지의 19년간 PSID 자료를 이용하여 표본이탈에 영향을 주는 요인을 살펴보았는데, 기혼자 보다는 미혼자일수록, 고령층일수록, 백인이 아닌 유색인종일수록, 고소득 가구일수록 그리고 소득의 변동폭이 클수록 이탈확률이 높은 것으로 나타났다. MaCurdy 등 (1998)은 NLSY(national longitudinal survey on youth) 자료를 이용하여 성별로 나누어 표본이탈 분석을 실시하였는데, 남성은 백인이면서 대졸 학력의 미취업자가 취업자에 비해 이탈률이 훨씬 높은 것으로 나타났다. 여성은 고졸과 대졸 학력을 가진 경우, 고소득일수록 이탈률이 높은 것으로 나타

났다. Zabel (1998)은 1968년부터 1988년까지 21년간의 PSID 자료와 1984년과 1990년 SIPP(survey of income and program participation) 자료를 이용하여 이탈모형을 세워 살펴보았는데, 많은 웨이브와 빈번한 조사 경험이 이탈 가능성을 높인다고 주장하였다. 그리고 모형에 포함되어 있는 조사원 및 조사과정 관련 변수 대부분이 유의한 영향을 주는 것으로 나타났다.

국내에서 패널탈락과 관련한 연구로는 김대일 등 (2000)이 있는데, 1993년부터 1997년까지의 경제활동인구조사를 패널화한 자료와 대우패널을 분석하였다. 그 결과, 성별로는 남성이, 연령은 고연령층이, 혼인상태는 이혼 또는 미혼이, 학력은 고학력자가, 고용형태별로는 임금근로자가, 가구주가, 경제활동상태별로는 실업자 또는 미취업자가 탈락을 많이 하는 것으로 나타났다. 이상호 (2003)는 한국고용정보원의 청년패널(youth panel) 자료를 이용하여 표본이탈 요인분석을 실시하였는데, 응답자 전체뿐만 아니라 노동시장 미진입자와 노동시장 진입자로 대상을 세분화하여 로짓모형을 통해 이탈여부에 영향을 주는 요인을 탐색하였다. 탐색 결과, 가구 소득, 경제활동상태, 자가소유여부 등 사회경제적 특성들이 패널탈락에 영향을 주는 것으로 나타났다. 또한 노동시장 진입여부에 따라 성별 응답확률의 차이가 있는 것을 찾아냈으며, 다른 연구들과 달리 소득이 증가할수록 응답확률이 높아지는 것으로 나타났는데, 그 이유는 조사대상자가 다른 연구들과 달리 청년층에 해당하는 15세에서 29세 사이로 국한했기 때문인 것으로 생각한다. 이상호 (2005)는 한국노동연구원의 한국노동패널(KLIPS; korean labor and income panel study) 자료를 이용하여 표본이탈에 영향을 주는 요인을 살펴보았는데, 프로빗 모형(probit model)을 이용하여 응답확률을 계산하고, 2단계 회귀모형을 통해 가구소득을 추정하였다. 결과에 따르면, 가구주는 연령이 높을수록 응답확률이 높게 나타났고, 미혼자, 배우자가 없는 기혼자, 배우자가 있는 기혼자의 순으로 응답확률이 높아지는 것으로 나타났으며, 성별변수는 유의한 효과를 주지 못하는 것으로 나타났다.

3. 자료소개 및 기초통계량

3.1. 대졸자 직업이동 경로조사 자료 소개 및 조사방법

본 연구에서 사용한 자료는 한국고용정보원의 GOMS 1차 년도에 해당하는 2006년 자료 (천영민 등, 2008)와 2차 년도에 해당하는 2007년 자료 (천영민 등, 2009)이다. GOMS는 한국고용정보원에서 2006년에 처음 실시한 패널조사이다. 강석훈과 김영원 (2008, 2009)에 따르면, 조사대상은 2004년 8월 및 2005년 2월에 전문대학 또는 대학을 졸업한 502,764명 중 학교유형, 권역(지역), 전공계열, 성별을 고려하여 표본추출된 26,544명이다. 학교유형은 전문대, 4년제, 교육대 등 3가지이고, 권역은 서울권, 경기권(인천, 경기, 강원), 충청권(대전, 충남, 충북), 경상권(부산, 대구, 울산, 경남, 경북), 전라권(광주, 전남, 전북, 제주) 등 5가지이고, 전공계열은 인문, 사회, 교육, 공학, 자연, 의약, 예체능 등 7가지로 구성되어 있다. 표 3.1의 2차년도 조사결과에 따르면, 원표본 26,544명 중 23,594명(88.9%)이 응답하였는데, 해외유학·군입대·사망 등의 이유로 조사 불가능한 경우가 원표본의 1.3%(347명)를 차지하여, 조사불가자를 제외한 최종 표본유지율은 90.1%로 나타났다. 조사불가 사유를 보면, 해외 84.7%(유학 30.5%, 장기해외거주 35.2%, 해외출장 및 해외취업 11% 등), 군입대 9.5%, 입원 4.6%, 사망 1.2%(4명)의 비율로 구성되어 있다. 표 3.1은 GOMS의 2차년도 패널 구축결과를 층화변인별로 나타낸 것인데, 층화변인별 표본유지율은 큰 차이를 보이지 않고 있다. 단, 표 3.1에서 조사가능패널수는 원표본 중에서 군입대, 해외유학 등에 해당하는 조사불가자를 제외한 수치이다.

대부분의 패널조사에서 표본유지율의 하락폭이 가장 크게 나타나는 경우는 2차년도이다. 3차년도부터는 하락폭이 눈에 띄게 감소하게 되는데, 국내에서 실시된 패널조사들의 2차년도 표본유지율을 나타낸 것이 표 3.2이다. 조사 대상이 개인이냐 가구이냐에 따라 표본유지율은 달라질 것이고, 대상이 청소년과

표 3.1. 2차년도 GOMS 패널 구축 결과

	구분	모집단	원표본	조사가능 패널수	2차년도 응답자	표본유지율
	전체	502,764	26,544	26,197	23,594	90.1
대학유형	전문대	228,336	9,981	9,864	8,746	88.7
	4년제	268,833	15,910	15,697	14,234	90.7
	교육대	5,595	653	636	614	96.5
성별	남자	244,069	14,216	14,047	12,768	90.9
	여자	258,695	12,328	12,150	10,826	89.1
학교소재지	서울권	90,885	6,053	5,959	5,391	90.5
	경기권	127,873	6,594	6,488	5,782	89.1
	충청권	72,096	3,733	3,676	3,343	90.9
	경상권	141,412	6,600	6,539	5,907	90.3
	전라권	70,498	3,564	3,535	3,171	89.7
전공계열	인문	51,417	2,553	2,502	2,221	88.8
	사회	124,502	6,546	6,466	5,812	89.9
	교육	30,453	1,957	1,924	1,782	92.6
	공학	143,768	8,151	8,049	7,282	90.5
	자연	53,618	2,991	2,965	2,682	90.5
	의약	33,981	1,577	1,560	1,399	89.7
	예체능	65,025	2,769	2,731	2,416	88.5

표 3.2. 국내 패널조사의 2차년도 표본유지율 비교

구분	한국 노동패널 (KLIPS)	청년 패널 (YP2001)	한국교육·고용패널 (KEEP)	한국 청소년패널 (KYPS) 중2	한국 청소년패널 (KYPS) 초4	대졸자 직업 이동 경로조사 (GOMS)
조사단위	가구	개인	개인(학교)	개인(학교)	개인(학교)	개인
시작년도	1999	2001	2004	2003	2004	2006
조사대상	전국 가구	만 15~29세 청년층	중3, 일반계고3 실업계고3	중2	초4	2005년 2월 대졸자
원표본	5,000가구 (13,321명)	8,296명	6,000명	3,449명	2,844명	26,544명
2차년도 조사가능표본	5,000가구 (13,321명)	8,296명	5,817명	3,417명	2,815명	26,197명
2차년도 응답표본	4,379가구 (11,237명)	5,956명	5,256명	3,188명	2,707명	23,594명
2차년도 표본유지율	87.6% (84.4%)	71.8%	90.4%	93.3%	96.2%	90.1%

같이 이동이 적으냐 청년층과 같이 이동이 많으냐에 따라 달라질 것이다. 국내에서 실시되고 있는 패널조사 중 2차 이상 실시된 패널 조사의 2차년도 표본유지율을 보면, GOMS가 개인단위 패널조사임에도 불구하고 매우 높은 수준임을 알 수 있다. 청년패널의 경우에는 패널탈락이 가장 많은 1~2차년도 조사기간 동안의 실사방법에 큰 변화가 있었기 때문에 그 영향으로 인해 초기 패널탈락이 많이 이루어진 것으로 알려져 있다.

1차년도와 2차년도의 주된 조사방법은 설문지에 면접원이 기입하는 형태의 대면면접조사이다. 하지만 실제로 패널탈락을 최소화하여 패널유지율을 제고하기 위해서 그리고 3차년도의 조사방식으로 진행예정인 CAPI 조사준비의 일환으로 WEB 조사를 2차년도에 일부 병행하여 실시하였다.

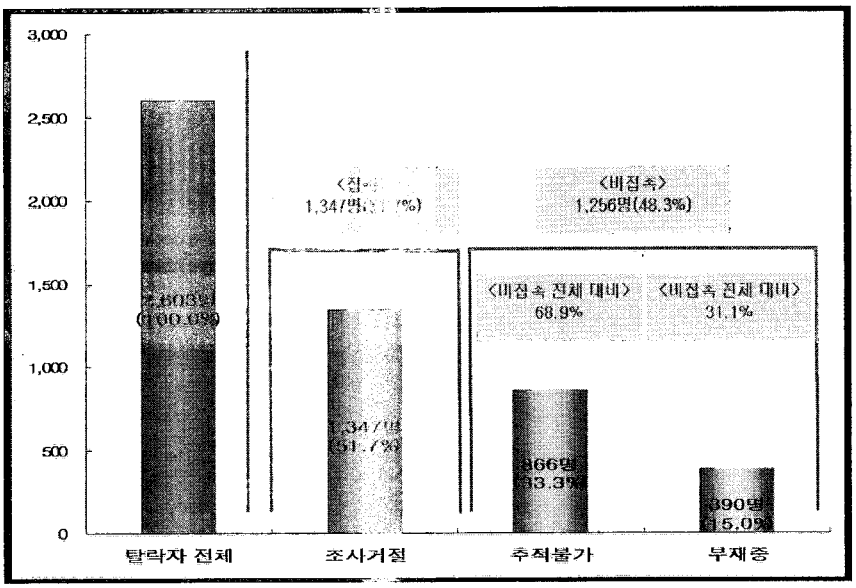

그림 **3.1.** 2차년도 패널탈락 유형

표 **3.3.** 패널 탈락자의 2차년도 경제활동상태

	구분	탈락자	탈락자내 비율	응답자	응답자내 비율
	전체	2,603	100.0	23,594	100.0
경제활동인구	취업자	1,964	75.5	19,042	80.7
	구직자	222	8.5	1,523	6.5
비경제활동인구	학생	144	5.5	1,282	5.4
	비학생	273	10.5	1,747	7.4

3.2. 패널탈락자 특성

그림 3.1에서 2차년도 조사 비성공자 2,603명의 패널탈락 유형을 보면, 추적불가 및 부재중으로 대상자를 접촉하지 못한 경우가 48.3%에 해당하는 1,256명이었으며, 대상자와 연락은 되었으나 개인적인 이유로 조사를 거절한 경우가 51.7%에 해당하는 1,347명이었다. 조사 거절자(접촉) 1,347명의 조사거절 사유를 살펴보면, '특별한 이유가 없음'과 '바쁘고 귀찮아서'가 78.5%로 대부분을 차지하였으며, 개인정보 노출 우려로 인한 조사거절은 0.7%에 불과한 것으로 나타났다. 거절 사유에서 알 수 있듯이, 피치 못 할 개인적인 사정이나 여건보다는 단순히 조사에 응하기 싫은 개인 성향에 따른 요인들이 대부분으로 적절한 유인책 마련을 통해 향후 조사에서 탈락자를 줄일 수 있음을 시사하고 있다.

2차년도 탈락자 특성을 자세히 살펴보면, 2차년도 조사 성공자(응답자)의 평균 연령은 26.7세, 2차년도 패널 탈락자의 평균연령은 26.4세로 큰 차이를 보이지 않는다. 또한 표 3.3에서 보는 바와 같이, 패널 탈락자의 1차 조사 응답당시 경제활동상태를 보면, 구직자와 학생을 제외한 비경제활동인구의 비율이 조사 성공자에 비해 조금 높게 나타나고 있다. 탈락자 중에서 구직자와 학생을 제외한 비경제활동인구 495명의 패널탈락 사유를 살펴보면, 89.1%가 대상자와의 접촉(추적불가 + 부재중)이 안 되어 조사에 실패한 경우였다. 탈락자 전체적으로는 응답 거절이 가장 많았으나, 구직자와 비경활은 본인 거취의 불안정으로 인해 연락이 안 되는 경우가 더 많았다.

표 4.1. 사용된 변수 설명

변수	변수의 값 설명
2차년도 조사 응답여부(종속변수)	탈락 = 0, 응답 = 1
혼인 상태	미혼 남성 여부, 미혼 여성 여부, 기혼 여성 여부
연령	25세 미만 여부, 30세 이상 여부
출신학교 유형	전문대 여부, 4년제 여부
비수도권대학 출신	비수도권출신 여부
경제활동상태	취업자 여부, 구직자 여부, 비경활 중 학생 여부
부모님과 동거	동거 여부
현재가구 월평균 총소득	100만~300만원 여부, 300만~500만원 여부, 500만원 이상 여부
동일면접원 여부(INTER1)	동일면접원 아님 = 0, 동일면접원임 = 1

4. 사용변수 및 분석방법

4.1. 사용변수

표 4.1은 분석에 사용된 변수에 대한 설명이다. 먼저 종속변수는 1차년도 조사 응답자의 2차년도 조사 성공여부로써, 응답하지 않은 패널탈락은 '0'으로 놓고, 조사성공한 응답은 '1'로 설정하였다. 설명변수에 대한 각 변수의 원래 범주값의 설정은 다음과 같다. 먼저 혼인상태는 기혼남성을 기준으로 하여 '0', 미혼 남성을 '1', 미혼 여성을 '2', 기혼 여성을 '3'으로 설정하였고, 연령은 25세 이상 30세 미만을 기준으로 하여 '0', 25세 미만을 '1', 30세 이상을 '2'로 설정하였고, 출신학교유형은 전문대를 '0', 4년제 대학을 '1', 교육대를 '2'로 설정하였다. 비수도권대학 출신 여부는 원래 5개의 범주로 구성된 권역 변수 대신에 사용되었는데, 출신대학이 비수도권이면 '0', 수도권이면 '1'로 설정하였다. 경제활동상태는 비경제활동인구 중에서 학생을 제외한 대졸자를 기준으로 하여 '0', 취업자를 '1', 미취업자를 '2', 비경제활동인구 중 학생을 '3'으로 설정하였고, 부모와의 동거여부는 동거하지 않음을 '0', 동거하고 있음을 '1'로 설정하였으며, 현재 가구의 월평균 총소득은 100만원 미만을 '0', 100~300만원을 '1', 300~500만원을 '2', 500만원 이상을 '3'으로 설정하였다. 하지만 실제 분석을 진행하기 위해 표 4.1에서 보는 바와 같이 범주형 자료를 모두 더미(dummy)로 처리하여 '0'과 '1'값만 갖도록 변환하여 사용하였다.

4.2. 분석방법

패널탈락의 요인으로는 크게 세 가지로, ① 성별, 혼인상태, 교육년수와 같은 인구통계학적 특성 및 가구소득이나 개인의 노동시장에서의 지위가 탈락에 미치는 영향, ② 이혼이나 실업 등과 같은 사회·경제적 충격이 응답 여부에 미치는 영향, ③ 응답자의 특성뿐만 아니라 면접원의 특성이나 조사시스템이 미치는 영향 등이다. 하지만, 2차년도 탈락자의 경제활동상태와 변화된 가족관계 등에 대한 정보 획득이 불가능하여 사회·경제적 충격에 의한 요인은 패널탈락 모형에 반영할 수 없었다. 따라서 본 연구에서는 패널탈락에 영향을 미치는 세 가지 요인 중 분석에 이용 가능한 일부 변인(① 인구통계학적 특성, ② 가구소득 및 경제활동상태, ③ 동일 면접원 여부)만을 가지고 패널탈락의 결정요인을 파악하였다. 본 논문에서 선택한 실증분석 방법으로는, 패널탈락여부를 종속변수(응답 = 1, 탈락 = 0)로 한 로짓(logit) 모형이다. 로짓 모형을 추정하기 위한 식은 다음과 같이 표현되며, 최우추정법(maximum likelihood estimation)으로 계수벡터를 추정하였다.

$$(1): \ L_i = \ln\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta X_i + u_i$$

$$(2): \ L_i = \ln\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta X_i + \gamma Z_i + u_i$$

$$(3): \ L_i = \ln\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta X_i + \gamma Z_i + \delta R_i + u_i$$

$$(4): \ L_i = \ln\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta X_i + \gamma Z_i + \delta R_i + \zeta Q_i + u_i$$

또한, 추가변수 Z, R, Q가 패널탈락 결정에 미치는 효과를 보기 위해, $H_0 : \gamma = 0 (H_1 : \gamma \neq 0)$(가설1), $H_0 : \delta = 0 (H_1 : \delta \neq 0)$(가설2), $H_0 : \zeta = 0 (H_1 : \zeta \neq 0)$(가설3)의 귀무가설(null hypothesis) 및 대립가설(alternative hypothesis)을 설정한 후, 제약된 모형과 제약되지 않은 모형의 최우추정(maximum likelihood estimation)에서 얻어진 로그최우도(log maximum likelihood)를 이용하여 $\lambda = -2(\ln Lr^* - \ln Lu^*) \sim \chi^2(r)$의 우도비검정통계량(likelihood ratio test statistics)을 계산하여 우도비검정을 실시하였다. 또한 의사결정나무(decision tree) 기법을 이용하여 패널탈락에 영향을 주는 요인을 탐색하였는데, 최적분리기준으로 카이스퀘어 검정(χ^2-test) 통계량을 사용하는 CHAID(chi square AID; Kass, 1980)를 사용하였다. 의사결정나무 알고리즘은 CHAID 외에도 Sonquist와 Morgan (1964)의 AID(automatic interaction detection), 분리기준으로 지니계수를 사용하는 Breiman 등 (1984)의 CART(Classification and Regression Trees), Quinlan (1983)이 제안한 ID3, 분리기준으로 엔트로피(entropy)를 사용하는 Quinlan (1993)의 C4.5 등이 있다. 일반적으로 의사결정나무에서 하나의 마디에 너무 많은 관측치가 있거나 너무 적은 관측치가 있을 경우에 오류가 증가하기 때문에 종료규칙을 설정하게 되는데, 본 연구에서 설정한 종료규칙은 다음과 같다.

1. 분리된 마디는 최소한 10개 이상의 관측치를 가져야 한다.
2. 마디의 관측치가 100개 이하일 경우에는 더 이상 분리하지 않는다.
3. 나무의 깊이(depth of tree)의 최대값은 6으로 설정하였다.
4. χ^2 검정시 값이 0.05이상일 경우에는 더 이상 분리하지 않는다.

본 연구에서 로짓 모형을 위해 사용된 통계분석 툴(tool)은 SAS 9.1.3이고, 의사결정나무 모형을 위해 사용된 프로그램은 SAS Enterprise Miner이다.

5. 분석 결과

5.1. 로짓 모형을 이용한 패널탈락 영향요인 분석결과

모형 (1)~(3)까지에 대한 로짓분석 결과를 표 5.1에 나타내고 있다. 우선 기혼 남성, 교육대 졸업자, 취업 상태, 부모와 동거, 가구소득이 높을수록 패널 유지율이 통계적으로 유의하게 높음을 알 수 있다. 우도비검정 결과, 모형 (2)와 모형 (3)은 모두 귀무가설을 기각하는 결과를 얻었다. 모형 (3)에서 출신대학의 소재지를 제외한 모든 설명변수가 통계적으로 유의미하게 나타나며, 모형 (1)의 추정과 비교하여 경제활동상태와 가구소득 변수를 추가한 경우에도 다른 설명변수의 유의성은 변화가 없는 것으로 나타났다. 기혼남성의 패널 유지율이 가장 높게 나타나며, 여성이든 남성이든 미혼자가 기혼자보다 탈락할 확률이 높은 것으로 보인다. 연령대의 경우 기준 변수인 20대 후반과 비교할 때 30대 이상의 통계적 유의성은 관찰되지 않으나, 20대 초반인 경우 20대 후반보다 유지율이 15% 더 높은 것으로 나타났다. 출신 대학 유형별로는 전문대보다 4년제 대학졸업자가 유지율이 약 1.3배, 교육대 졸업자는 3.7배나 높은 것으로 나타났다. 경제활동상태별로는 1차 조사당시 취업자인 경우의 유지율이 가장 높으며, 구직자의 경우 통계적으로 유의하지는 않지만 학생이 아닌 비 경제활동인구와 비슷한 유지율을 보이는 것으로 분

표 **5.1.** 로짓모형 추정치 - 개인특성

	모형 (1)		모형 (2)		모형 (3)	
	추정치	Exp(B)	추정치	Exp(B)	추정치	Exp(B)
상수	2.373(0.101)***	10.729	1.981(0.123)***	7.247	1.763(0.138)***	5.827
혼인상태						
미혼남성	−0.305(0.098)***	0.737	−0.267(0.098)***	0.766	−0.312(0.100)***	0.732
미혼여성	−0.567(0.102)***	0.567	−0.536(0.103)***	0.585	−0.607(0.106)***	0.545
기혼여성	−0.338(0.124)***	0.713	−0.220(0.126)*	0.803	−0.231(0.126)*	0.794
연령						
25세 미만	0.124(0.058)**	1.132	0.138(0.058)**	1.148	0.139(0.058)**	1.150
30세 이상	0.025(0.086)	1.026	−0.000(0.086)	1.000	0.015(0.086)	1.015
출신학교유형						
4년제	0.260(0.045)***	1.297	0.276(0.045)***	1.318	0.294(0.046)***	1.342
교육대	1.358(0.220)***	3.890	1.305(0.220)***	3.687	1.312(0.220)***	3.713
비수도권대학출신	0.027(0.042)	1.028	0.035(0.042)	1.036	0.058(0.043)	1.060
경제활동상태						
취업자			0.408(0.071)***	1.504	0.387(0.073)***	1.472
구직자			0.078(0.098)	1.081	0.082(0.098)	1.085
비경활중 학생			0.301(0.111)***	1.351	0.341(0.111)***	1.406
부모님과 동거					0.130(0.050)***	1.139
가구소득						
100~300만					0.163(0.080)**	1.177
300~500만					0.248(0.088)***	1.281
500만원 이상					0.230(0.098)**	1.259
N	26,080		26,080		26,080	
$-2\log L$	16777.507		16732.899		16707.024	
모형적합도	117.215***		161.823***		187.698***	
올바른 예측률	90.1%		90.1%		90.1%	
우도비검정			44.608***		25.874***	

석되었다. 마지막으로, 1차 조사당시 부모와 함께 살고 있었던 경우에 그렇지 않은 경우보다 유지율이 14% 더 높게 나타났으며, 가구소득이 300~500만원 미만인 경우의 유지율이 100만원 미만인 경우에 비해 약 1.3배로 가장 높으며, 그 다음으로 500만원 이상, 100~300만원 미만의 순으로 나타났다. 단, 표 5.1에서 모형적합도는 자체모형의 적합도를 검정하는 우도비검정통계량, 올바른 예측률은 패널탈락 결정에서 실제값과 예측값이 동일한 표본의 비중, 우도비 검정은 경제활동상태와 가구소득 변수의 유의성을 검정하는 우도비검정통계량을 의미하고 추정치의 괄호안의 숫자는 표준오차를 의미하며, * 표시는 각각 *: $p < .1$, **: $p < .05$, ***: $p < .01$에서 통계적으로 유의미한 것을 나타낸다.

면접원 유형에 따른 패널탈락의 효과를 보기 위해, 웹조사 응답자를 제외한 표본을 대상으로 로짓분석을 실시한 결과는 표 5.2에 나타나 있다. 모형 (3-1)의 경우, 모형 (3)의 추정과 비교하여 출신대학 소재지 변수를 제외하고 다른 설명변수의 유의성은 변화가 없는 것으로 나타났다. 또한, 가구소득이 높을수록 유지율이 높아지는 것으로 나타나, 모형 (3)과 약간의 차이를 보이고 있다. 우도비검정 결과, 모형 (4)의 경우도 귀무가설을 기각하는 결과를 얻었다. 즉, 패널탈락에 영향을 미치는 다른 요인들을 통제한 후에도 동일 면접원 여부가 조사에 응답하는데 유의한 효과를 미쳤다는 것을 알 수 있다. 모형 (4)에서 동일 면접원이 조사를 시도한 경우에 그렇지 않은 경우보다 1.5배나 높은 유지율을 보이는 것으로 나타났다. 표 5.2에서 우도비 검정은 면접원 유형 변수의 유의성을 검정하는 우도비검정통계량을 의미한다.

표 5.2. 로짓모형 추정치 - 개인특성 + 면접원유형(웹조사 제외)

	모형 (3-1)		모형 (4)	
	추정치	Exp(B)	추정치	Exp(B)
상수	1.644(0.139)***	5.176	1.549(0.139)***	4.707
혼인상태				
미혼남성	−0.321(0.100)***	0.726	−0.308(0.100)***	0.735
미혼여성	−0.636(0.106)***	0.530	−0.619(0.106)***	0.538
기혼여성	−0.218(0.127)*	0.804	−0.227(0.127)*	0.797
연령				
25세 미만	0.118(0.058)**	1.125	0.113(0.058)*	1.119
30세 이상	0.031(0.086)	1.032	0.019(0.086)	1.020
출신학교유형				
4년제	−0.207(0.046)***	1.230	0.216(0.046)***	1.241
교육대	−1.246(0.221)***	3.478	1.245(0.221)***	3.472
비수도권대학출신	0.074(0.043)*	1.076	0.038(0.043)	1.038
경제활동상태				
취업자	−0.409(0.074)***	1.506	0.411(0.074)***	1.508
구직자	0.106(0.099)	1.112	0.121(0.099)	1.128
비경활중 학생	−0.330(0.113)***	1.391	0.318(0.113)***	1.474
부모님과 동거	−0.140(0.051)***	1.150	0.128(0.051)**	1.137
가구소득				
100~300만	0.170(0.081)**	1.186	0.159(0.081)**	1.172
300~500만	−0.253(0.089)***	1.288	0.240(0.089)***	1.271
500만원 이상	−0.268(0.098)***	1.294	0.250(0.099)**	1.284
동일면접원 조사시도			0.438(0.050)***	1.549
N	22,851		22,851	
$-2\log L$	15985.820		15904.570	
모형적합도	183.911***		265.161***	
올바른 예측률	88.6%		88.6%	
우도비검정			81.250***	
Max-rescaled R^2	0.0158		0.0227	

5.2. 의사결정나무를 이용한 패널탈락영향요인 분석 결과

앞에서 언급한 바와 같이, 본 연구에서는 의사결정나무기법 중에서 CHAID를 이용하였는데, 목표변수(target variable)는 패널탈락여부이다. 입력변수(input variable)는 앞의 로짓모형에서 사용한 변수와 동일한 변수를 이용하였다. 모든 변수는 이진값(binary value) 형태로 만들기 위해 더미 처리하여 사용하였고, 이진분리(binary split)를 이용하여 의사결정나무모형을 구성하였다. 총 22,851명의 자료 중에서 67%에 해당하는 15,310명의 자료를 훈련데이터(training data)로 사용하여 의사결정나무모형을 생성하는 데 사용하였다. 그리고 나머지 33%에 해당하는 7,541명의 자료를 유효성 검증데이터(validation data)로 사용하여, 훈련데이터로 만든 모형을 평가하는 데 사용하였다. 본 연구에서 작성한 최종모형은 그림 5.1과 같이 6개의 잎을 가진 의사결정나무이다. 뿌리마디에서 제일 먼저 분리기준으로 선택된 변수는 동일면접원여부(INTER1)이다. 두 번째 마디에서 분리기준으로 선택된 변수는 취업자여부이며, 세 번째 분리기준 변수는 서로 다른데, 4년제 대학 졸업여부와 교육대 졸업여부이다. 앞의 로짓모형 결과에서도 동일면접원 조사시도가 유의한 영향을 준 것으로 나타났는데, 의사결정나무 모형에서도 가장 중요한 변수로 선택되어, 패널탈락여부는 응답자의 개인적 특성보다는 동일한 면접원이

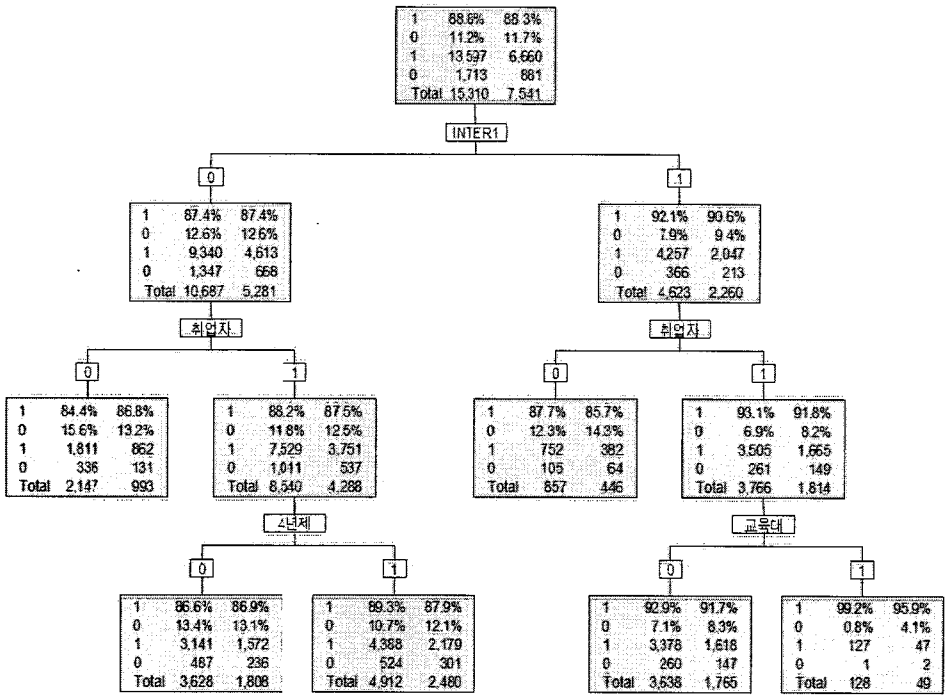

그림 **5.1.** 의사결정나무 모형

조사했느냐에 따라 표본유지가 결정되기 때문에 패널조사에서는 동일한 면접원 운영에 대해 고민할 필요가 있는 것으로 나타났다.

의사결정나무 모형을 통해 얻어진 의사결정규칙(decision rule)을 정리하면 다음과 같다.

1) 미취업자이고 동일면접원이 아니면, 2,147명의 훈련데이터 중에서 84.4%에 해당하는 1,811명은 조사 성공하였고, 15.6%에 해당하는 336명은 패널탈락한 것으로 나타났다.

2) 취업자이고 동일면접원이 조사하면, 857명의 훈련데이터 중에서 87.7%에 해당하는 752명은 조사 성공하였고, 12.3%에 해당하는 105명은 패널탈락한 것으로 나타났다.

3) 4년제 졸업자가 아니고 취업자이면서 동일면접원이 조사하지 않으면, 3,628명 중에서 86.6%에 해당하는 3,141명은 조사 성공하였고, 13.4%에 해당하는 487명은 패널탈락한 것으로 나타났다.

4) 4년제 대학교 졸업자이고 취업자이면서 동일면접원이 조사하지 않으면, 4,912명 중에서 89.3%에 해당하는 4,388명은 조사 성공하였고, 10.7%에 해당하는 524명은 패널탈락한 것으로 나타났다.

5) 교육대 졸업자가 아니고 취업자이면서 동일면접원이 조사하면, 3,638명 중에서 92.9%에 해당하는 3,378명은 조사 성공하였고, 7.1%에 해당하는 260명은 패널탈락한 것으로 나타났다.

6) 교육대 졸업자이고 취업자이면서 동일면접원이 조사하면, 128명의 99.2%에 해당하는 127명은 조사 성공하였고, 단지 0.8%에 해당하는 1명만이 패널탈락한 것으로 나타났다.

이상의 결과를 가지고 보면, 1)~3)의 경우는 조사 성공률이 낮으므로 향후 조사시 관심을 가져야 할 것이다. 6)의 경우는 거의 대부분의 응답자에 대한 조사가 성공한 것으로 나타났다. 한편 취업자에 비해 미취업자들이 탈락할 가능성이 높은 것을 유추해 볼 수 있다.

6. 결론

노동시장 미진입자, 미혼, 20대 후반, 부모와 비동거 등 개인 상태의 변동(이동) 가능성이 높은 응답자의 패널탈락 확률이 높음을 알 수 있었다. 하지만, 우도비(likelihood ratio) 테스트 결과, 전체 모형이 통계적으로 유의미한 설명력을 가지나, 모형의 전체적인 설명력은 2.3%에 불과하여 본 연구를 통해 패널탈락을 설명하기에는 부족한 점이 있었다. 즉, 패널탈락자가 특정 계층이나 경제활동상태에 집중되어 있지 않다고 해석할 수 있다. 따라서 패널탈락의 요인이 패널 개인의 인구통계학적 특성에 의한 것이라기보다는 조사시스템의 효과, 응답자 사례, 사회·경제적 분위기 등 나머지 97%에 의한 것으로 보인다. 그럼에도 불구하고 패널탈락에 가장 영향을 미치는 변수는 동일면접원 여부인 것으로 나타났다. 하지만 동일면접원 확보는 동일한 조사업체가 다시 조사용역을 수행해야 할 뿐만 아니라 전년도 동일 조사에 투입된 조사원이 재투입되어야 하는 등의 제약조건이 많기 때문에 쉽게 접근할 수 없는 한계를 갖게 된다.

선행연구 결과에서 보듯이, 자료의 대상에 따라 패널탈락에 미치는 요인이 달라지는 것을 알 수 있었다. 또한 면접원의 효과나 조사시스템의 문제 등이 더 큰 원인이 될 수도 있음을 알 수 있었다. 실제로 조사성공의 열쇠는 응답자에게 달려있기보다는 조사시스템에 달려 있다고 볼 수 있다. 조사시스템관련 변수 중에서 면접원과 관련한 것들이 중요한 영향을 줄 것으로 생각한다. 특히 조사경력년수, 패널조사 경험여부, 고용관련조사 경험여부, 응답자와 동일성별여부, 조사원의 연령 및 학력, 응답자와 동일거주지 여부 등을 추가적으로 확인할 필요가 있을 것으로 생각한다. 이를 바탕으로 하여, 조사 성공요인을 도출하여 향후 추적조사의 면접원 교육에 활용할 필요가 있다.

패널탈락의 문제는 조사의 장기적 성공을 위해서 매우 중요한 주제이다. 추가적인 연구를 통해 체계적인 조사시스템 구축, 응답자 관리 방법 개발뿐만 아니라 패널탈락의 편의를 보완하기 위한 가중치 부여 등 다양한 방법을 도입하여 GOMS 자료의 신뢰성을 확보해야 할 것이다. 뿐만 아니라 3차년도 조사에서부터는 탈락된 패널을 다시 복귀할 수 있도록 하는 방법에 대해서도 고민을 해야 할 것이다.

참고문헌

강석훈, 김영원 (2008). <2006년 대졸자 직업이동 경로조사 가중치 부여 방법 연구>, 한국고용정보원.

강석훈, 김영원 (2009). <2007년 대졸자 직업이동 경로조사 가중치 부여 방법 연구>, 한국고용정보원.

김대일, 남재량, 류근관 (2000). 한국노동패널 표본의 대표성과 패널조사 표본 이탈자의 특성 연구, <노동경제론집>, **23**, 1-33.

이상호 (2003). 청년패널의 표본이탈 요인분석, <제2회 산업·직업별 고용구조조사 및 청년패널 심포지엄>, 한국고용정보원.

이상호 (2005). 한국노동패널의 표본이탈 분석, <노동리뷰>, **11**, 66-80.

천영민, 박상현, 정승철, 윤정혜, 이성재 (2008). <2006 대졸자 직업이동 경로조사 기초분석 보고서>, 한국고용정보원.

천영민, 정승철, 윤정혜, 이성재, 이주현, 심재훈, 서지연 (2009). <2007 대졸자 직업이동 경로조사 기초분석 보고서>, 한국고용정보원. <고려대 경상논집>, 1호, 321-329.

Becketti, S., William, G., Lee, L. and Finis, W. (1998). The panel study of income dynamics after fourteen years: An evaluation, *Journal of Labor Economics*, **6**, 472-492.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, Chapman & Hall, New York.

Fitzgerald, J., Gottschalk, P. and Moffit, R. (1998). An analysis of sample attrition in panel data, *The Journal of Human Resources*, **33**, 251-299.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of data, *Applied Statistics*, **29**, 119-127.

Lillard, L. A. and Panis, C. W. A. (1998). Panel attrition from the PSID. *The Journal of Human Resources*, **33**, 437-457.

MaCurdy, T., Mroz, T. and Gritz, R. M. (1998). An evaluation of the national longitudinal survey on youth, *The Journal of Human Resources*, **33**, 345–436.

Quinlan, J. R. (1983). *Learning Efficient Classification Procedures.* Machine Learning: An Artificial Intelligence Approach, Palo Alto, CA:Tioga Press.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA.

Sonquist, J. A. and Morgan, J. N. (1964). *The Detection of Interaction Effects.* Ann Arbor: Institute for Social Research, University of Michigan.

Zabel, J. E. (1998). An analysis of attrition in the panel study of income dynamics and the survey of income and program participation with an application to a model of labor, *The Journal of Human Resources*, **33**, 479–506.

An Analysis of Panel Attrition in GOMS(Graduates Occupational Survey)

Young-Min Chun[1] · Jeong-Hye Yoon[2] · Min-Hong Oh[3]

[1]Center for Employment Survey and Analysis, Korea Employment Information Service
[2]Center for Employment Survey and Analysis, Korea Employment Information Service
[3] Center for Employment Projection, Korea Employment Information Service

(Received July 2009; accepted August 2009)

Abstract

It would cause a serious problem in the panel data when panel attrition is concentrated on certain socio-economic groups. Using the GOMS, this study investigates whether there exists non-random attrition bias in the data and seeks for feasible solutions to minimize the bias. The results of logit analyses show that panel attrition in the GOMS results mainly from surveying system but not from the surveyed. Therefore, the result suggests to develop well-organized management skill and systems as well as to construct weighting methods.

Keywords: CAPI, decision tree, logit model, panel attrition.

[3]Corresponding author: Research Fellow, Center for Employment Projection, Korea Employment Information Service, 77-11, Mullae-dong 3-ga, Yeongdeungpo-gu, Seoul, 150-093, Korea.
E-mail: mhoh7266@keis.or.kr