연구노트

한국청소년패널조사(KYPS) 가중치 부여 방법 연구:
중학교 2학년 패널의 경우*

A Study on the Construction of Weights for KYPS

박민규** · 이경상*** · 박현수**** · 강현철*****
Mingue Park · Kyeong−Sang Lee · Hyun−Soo Park · Hyuncheol Kang

본 연구에서는 2003년 시작된 한국청소년패널조사 중학교 2학년 패널자료 분석을 위해 필수적인 각 연도별 횡단면, 종단면 가중치 산출방안을 소개하고 있다. 패널 모집 당시 추출된 표본이 대표하는 모집단의 종단면적 변화 분석을 위한 종단면 가중치는 로지스틱 회귀분석을 이용한 무응답 보정과 사후 층화를 통해 산출되었으며, 조사 연도의 표본 대응 모집단 분석을 위해 필수적인 횡단면 가중치는 전년도 대비 응답률과 사후 층화를 통해 산출되었다.

주제어: 패널조사, 종단면 가중치, 횡단면 가중치, 응답률, 로지스틱 회귀분석

We introduced the methodologies used to construct the longitudinal weights and cross−sectional weight that are required for the analysis of Korea Youth Panel Survey. To analyze the longitudinal dynamic change of the population, we derived the longitudinal weight through nonresponse adjustment based on logistic regression and post−stratification. Cross−sectional weights that are necessary to produce an asymptotically unbiased estimator of the population parameter were constructed through simple nonresponse adjustment based on overall response rate and

* 이 논문은 2009년 한국청소년정책연구원 '한국청소년패널조사(KYPS)' 연구의 일부분을 수정·보완한 것임.

** 고려대학교 통계학과 부교수

*** 한국청소년정책연구원 연구위원

**** 충북대학교 사회학과 강사

***** 교신저자(corresponding author): 호서대학교 정보통계학과 부교수 강현철.
E−mail: hychkang@hoseo.edu.

post-stratification.

key words: panel survey, longitudinal weights, cross-sectional weight, response rate, logistic regression

Ⅰ. 서 론

한국청소년패널조사(KYPS: Korea Youth Panel Survey)는 2003년 청년실업의 여파, 교실붕괴와 학업중단의 증가 및 일탈행위 등으로 인해 밝은 미래를 설계하는 데 많은 어려움을 겪고 있는 우리 청소년들에게 적절한 사회문화적 지원을 제시하고자 잠재적 직업선택, 향후 진로설정 및 준비, 일탈행위, 여가참여 등의 횡단적 실태 및 종단적 변화양상 및 그 원인에 대해 파악하고자 하는 목적에서 시작되었다(이경상 · 안선영 2009; 한국청소년정책연구원 2009).

이를 위해 2003년 기준 전국 중학교 2학년 청소년들 중 대표성 있게 표집된 3,449명의 청소년들 및 부모들(2003년 조사 시작~2008년까지 추적조사)과 2004년 기준 전국 초등학교 4학년 청소년들 중 대표성 있게 표집된 2,844명의 청소년들 및 부모들(2004년 조사 시작~2008년까지 추적조사)을 대상으로 하여 진로, 일탈, 여가 등의 생활실태에 대해 전망적 패널조사(prospective panel survey)의 방법으로 동일표본을 반복적으로 추적·조사하여 종단적 패널 데이터를 구축하고자 하였다. 또한 종단적 패널 데이터의 활용을 극대화하고 연구성과로 축적시키기 위해 구축된 횡단적·종단적 데이터를 관련 학계의 교수, 연구원, 대학원생 등에게 개방함으로써 이 데이터를 활용한 학술논문을 축적시키고자 하였다(김경동 · 이온죽 1993; 김영석 1999; 이경상 · 안선영 2009; 한국청소년정책연구원 2009; Babbie 2001).

2008년 마지막 추적조사 결과 청소년 기준으로 중학교 2학년 패널의 경우 82.1%의 표본을 유지하였고, 초등학교 4학년 패널의 경우 86.1%의 표본을 유지하였다. 또한 구축된 데이터를 관련 학계에 제공하여 2011년 현재까지 한국청소년패널조사 데이터를 활용한 논문이 400여 편 축적되었다.

패널자료의 분석을 위해서는 각 조사 시점별 횡단면 분석과 또한 조사 시점까지의 종단면 분석을 위한 횡, 종단면 가중치의 산출이 필수적이다. 본 연구에서는 이용한 한국

청소년패널조사 데이터의 분석을 위해 제공된 횡, 종단면 가중치의 산출과정을 원년도 가중치와 3차년도 종단면 가중치 산출내용을 중심으로 설명하고자 한다. 가중치 산출을 위해서는 기본적으로 로지스틱 회귀분석과 사후층화 방법이 사용되었으며 이에 대한 전반적인 내용은 Groves et al.(2002)와 Sarndal & Lundstrom(2006)에 연구되어 있다.

Ⅱ. 1차년도 가중치 부여 방법

한국청소년패널조사를 위한 1차년도 표본추출을 위해서 층화 2단계 집락추출법이 사용되었다. 각 중학교의 층화를 위해서는 12개(7개 광역시, 경기도, 강원도, 충청도, 전라도, 경상도)의 지역 층이 사용되었으며 1차 및 2차 추출단위로는 각각 중학교와 중학교 내의 학급이 고려되었다. 즉, 제 1단계에서는 각 층 내에서 중학교가 추출되었으며 이어 2단계에서는 추출된 중학교 내의 2학년 학급들 중 1개의 학급이 추출되었다. 추출된 학급 내의 모든 학생들은 본 조사를 위한 패널로 모두 선택되었다.

〈표 1〉 12개 시도별 표집학교 수

(단위: 명)

구 분	중학교 2학년 학생 수	중학교 수	중학교 2학년 학급 수	학급당 평균 학생 수	표집 학교 수
총 계	618,100	2,808	17,504	35	104
서 울	120,826	358	3,591	34	20
부 산	48,040	165	1,374	35	8
대 구	36,529	111	977	37	6
인 천	36,596	103	912	40	6
광 주	20,529	71	533	39	4
대 전	20,227	73	570	35	3
울 산	16,478	46	432	38	3
경 기	135,084	434	3,436	39	21
강 원	19,117	159	573	33	4
충 청	42,286	310	1,324	32	8
전 라	48,482	448	1,582	31	8
경 상	73,906	530	2,200	34	13

1차 추출단위인 중학교 추출을 위한 표집틀로는 2003년 4월 1일 기준 교육통계연보가 사용되었으며 표집틀로부터의 중학교 추출을 위해서는 각 중학교의 2학년 학생 수를 크기(size) 변수로 사용한 확률비례추출법(probability proportional to size: PPS)이 사용되었다. 단 실사가 불가능한 도서지역과 중학교 2학년 학급당 평균 학생 수가 20명 미만인 중학교는 표집틀로부터 제외되었다. 각 층별 표본 중학교 수 결정은 지역별 중학교 2학년 학생 수 및 학급당 평균 학생 수를 이용하여 이루어졌다. <표 1>은 12개 광역시·도별 중학교 2학년 학생 수, 표집틀의 중학교 수, 학급당 평균 학생 수 그리고 최종 추출된 학교 수를 나타내고 있다.

1차년도 패널자료 분석을 위한 가중치는 크게 1) 표본 가중치 계산과 2) 사후층화의 두 단계를 거쳐 산출되었다. 1차년도의 경우, 종단면 분석이 이루어지지 않을 뿐 아니라 그 자료구조상의 이유로 횡단면 가중치만이 산출된다. 표본 가중치(sampling weight)는 관측단위인 중학생이 표본에 포함될 확률인 표본포함확률(inclusion probability)의 역수로 정의되며 이는 표본에 추출된 각 중학생이 대표하는 모집단에서의 중학생 수를 나타낸다. 언급된 표본 추출과정을 통해 추출된 h번째 층 내 i번째 학교의 j번째 학급에 속한 k번째 중학생에게 부여되는 표본 가중치는 다음과 같다.

$$\alpha_{hijk} = \left[\left(\frac{n_h \sum_j x_{hij}}{\sum_i \sum_j x_{hij}} \right) \left(\frac{1}{m_{hi}} \right) \right]^{-1} \quad (1)$$

여기서 x_{hij}는 h번째 층 내 i번째 학교의 j번째 학급의 학생 수, n_h는 h번째 층의 표본학교 수를 그리고 m_{hi}는 h번째 층 내 i번째 학교의 학급 수를 각각 나타낸다.

표본추출을 위해서 각 지역별 그리고 각 중학교의 학생 수를 감안한 표본추출 방안이 고려되었으나 중학생들의 인구학적 변인별 분포 특별히 성별, 지역별 모집단 분포가 표본추출과정에서 반영되지 않았다. 그러나 실제 조사가 이루어진 관심 변수들이 이러한 인구학적 변인들과 밀접한 관계가 있음으로 인하여, 1차년도 최종 가중치 산출을 위해서 성과 지역 변수의 모집단 분포를 이용한 사후층화가 이루어졌다. 사후층화를 통해 산출된 사후층 p에 속한 중학생의 최종 가중치는 다음과 같이 표현된다.

$$w_{hijk} = \alpha_{hijk} \left(\frac{N_p}{\widehat{N}_p} \right). \quad (2)$$

여기서 α_{hijk}는 표본가중치이고,

$$\boldsymbol{t}_z = \sum_U \boldsymbol{z}_{hijk},$$

$$\hat{\boldsymbol{t}}_z = \sum_S \alpha_{hijk}\boldsymbol{z}_{hijk},\ \boldsymbol{z}_{hijk} = (z_{1,hijk},\cdots,z_{P,hijk}),\ N_p = \sum_U z_{p,hijk},$$

$$\hat{N}_p = \sum_S \alpha_{hijk} z_{p,hijk},$$

그리고 $z_{p,hijk}$는 h번째 층 내 i번째 학교의 j번째 학급에 속한 k번째 중학생이 사후 층 p에 속하면 1, 그렇지 않은 경우 0의 값을 갖는 지시변수(indicator variable)이다.

식 (2)를 통하여 산출된 1차년도 횡단면 가중치의 요약 통계량 및 지역별·성별 분포가 〈표 2〉 및 〈표 3〉에 제시되어 있다. 〈표 2〉를 보면 여자 가중치의 평균이 남자 가중치의 평균보다 작음을 볼 수 있는 데, 이는 모집단 분포에 비하여 여자가 상대적으로 많이 표집된 상황이 반영된 것이다. 결론적으로 가중치 부여 이후의 지역별·성별 중학교 2학년 학생 수 추정량의 분포가 모집단 분포와 동일하게 되도록 조정되어 있다.

Ⅲ. t차년도 횡단면 가중치 부여 방법

패널 조사가 진행됨에 따라 무응답이 발생하게 되며 무응답의 패턴에 따라 매년 통계적으로 타당한 가중치 산출이 필수적이다. 한국청소년패널의 횡단면 분석을 위한 가중치 보정은 종단면 가중치 산출과정에 비하여 단순하게 이루어졌다. 종단면 가중치는 한 시점에서 조사에 응하지 않는 관측치의 경우 이 후 모든 시점에서 가중치가 산출되지 않는 반면 횡단면 가중치는 매 조사 시점 관측된 모든 관측치에 가중치가 부여된다. 즉 횡단면 가중치 부여 대상 관측치의 결측패턴이 단조적인 성격을 지니고 있지 않다. 또한 시간이 지남에 따라 종단면 가중치 산출을 위한 표본 감소 비율에 비하여 횡단면 가중치를 산출해야 하는 대상 관측치의 감소 비율 역시 낮게 나타나고 있다. 이러한 사항들을 고려하여 횡단면 가중치 산출을 위한 무응답 보정은 각 조사 시점별 응답률을 이용하여 일괄적으로 처리 후 사후층화를 통한 각 사후층별 보정을 실시하는 방법을 이용하였다. 구체적으로 t차년도 횡단면 가중치는 다음과 같은 절차를 수행하여 계산되었다. 먼저 기본가중치는 표본탈락률을 보정하기 위하여 전년도 대비 응답률의 역수를 전년도 횡단면 가중치에 곱하여 계산되었다. 즉,

〈표 2〉 1차년도 횡단면 가중치의 요약 통계량

(단위: 명)

	n	최소값	최대값	평균	합계	표준편차	변동계수
전체	3,449	90	318	179	618,100	37	21
남자	1,725	116	318	190	327,782	38	20
여자	1,724	90	292	168	290,317	33	20

〈표 3〉 1차년도 지역별·성별 표본 분포 및 추정량의 분포

(단위: 명)

	표본 분포						추정량 분포					
	남 자		여 자		전 체		남 자		여 자		전 체	
	n	%	n	%	n	%	N	%	N	%	N	%
전체	1,725	50.0	1,724	50.0	3,449	100.0	327,782	53.0	290,317	47.0	618,100	100.0
서울	296	49.7	299	50.3	595	17.3	64,097	53.0	56,728	47.0	120,826	19.5
부산	102	40.0	153	60.0	255	7.4	25,659	53.4	22,381	46.6	48,040	7.8
대구	114	53.8	98	46.2	212	6.1	20,442	56.0	16,087	44.0	36,529	5.9
인천	102	51.3	97	48.7	199	5.8	18,878	51.6	17,718	48.4	36,596	5.9
광주	79	57.7	58	42.3	137	4.0	10,734	52.3	9,795	47.7	20,529	3.3
대전	49	48.5	52	51.5	101	2.9	10,973	54.2	9,254	45.8	20,227	3.3
울산	52	49.1	54	50.9	106	3.1	9,156	55.6	7,322	44.4	16,478	2.7
경기	325	42.5	440	57.5	765	22.2	70,370	52.1	64,714	47.9	135,084	21.9
강원	66	53.2	58	46.8	124	3.6	9,932	52.0	9,185	48.0	19,117	3.1
충청	130	50.2	129	49.8	259	7.5	22,220	52.5	20,066	47.5	42,286	6.8
전라	142	58.2	102	41.8	244	7.1	25,157	51.9	23,325	48.1	48,482	7.8
경상	268	59.3	184	40.7	452	13.1	40,165	54.3	33,741	45.7	73,906	12.0

$$\alpha_{t,hijk}^{c} = w_{t-1,hijk}^{c}\,\hat{p}_{t,hijk}^{-1} \tag{3}$$

여기서 $w_{t-1,hijk}^{c}$는 $t-1$차년도 횡단면 가중치(혹은 최근 응답한 시점의 횡단면 가중치), $\hat{p}_{t,hijk}$는 $(t-1)$차년도 대비 응답률을 나타낸다. 예를 들어, 2차년도의 횡단면 기본 가중치는 1차년도 대비 응답률(3,188/3,449 = 92.4%)의 역수를 1차년도 횡단면 가중

치에 곱하여 계산되었다. 결론적으로 각 학생에 대한 2차년도의 횡단면 기본 가중치는 해당 학생의 1차년도 횡단면 가중치의 약 1.08(= 1/0.924)배가 된다.

다음으로 사후층화를 통한 최종 가중치가 산출되었다. 이는 t차년도 횡단면 조사 분석 결과가 당 해의 모집단에 대해 타당한 추론을 가능하게 하기 위함이다. 사후층화를 통해 산출된 사후층 p에 속한 중학생의 t차년도 최종 횡단면 가중치는 다음과 같이 표현된다.

$$w^{c}_{t,hijk} = \alpha^{c}_{t,hijk}\left(\frac{N_{t,p}}{\hat{N}_{t,p}}\right) \tag{4}$$

여기서 $N_{t,p} = \sum_{U} z_{t,p,hijk}$, $\hat{N}_{t,p} = \sum_{S} \alpha^{c}_{t,hijk} z_{t,p,hijk}$, 그리고 $z_{t,p,hijk}$는 h번째 층 내 i번째 학교의 j번째 학급에 속한 k번째 중학생이 t차년도에 사후층 p에 속하면 1, 그렇지 않은 경우 0의 값을 갖는 지시변수이다.

이러한 절차를 통하여 산출된 $t(= 2, \cdots, 6)$차년도 횡단면 가중치의 요약 통계량이 〈표 4〉에 제시되어 있으며, 2차년도 횡단면 가중치의 지역별·성별 분포가 〈표 5〉에 제시되어 있다.

〈표 4〉 t차년도 횡단면 가중치의 요약 통계량

(단위: 명)

t		n	최소값	최대값	평균	합계	표준편차	변동계수
2	전체	3,188	101	355	193	614,189	44	23
	남자	1,594	116	355	204	325,652	48	23
	여자	1,594	101	323	181	288,537	36	20
3	전체	3,125	100	349	196	611,770	44	22
	남자	1,572	120	349	207	325,442	45	22
	여자	1,553	100	343	184	286,328	39	21
4	전체	3,121	100	357	192	597,895	43	22
	남자	1,566	120	357	203	317,365	45	22
	여자	1,555	100	332	180	280,529	36	20
5	전체	2,967	104	376	199	589,673	47	23
	남자	1,510	119	376	207	311,988	49	24
	여자	1,457	104	371	191	277,684	42	22
6	전체	2,833	109	387	217	614,415	45	21
	남자	1,348	141	387	230	310,423	47	20
	여자	1,485	109	356	205	303,992	40	20

〈표 5〉 2차년도 횡단면 가중치의 지역별/성별 분포

(단위: 명)

	표본 분포						추정량 분포					
	남 자		여 자		전 체		남 자		여 자		전 체	
	n	%	n	%	n	%	N	%	N	%	N	%
전체	1,594	50.0	1,594	50.0	3,188	100.0	325,652	53.0	288,537	47.0	614,189	100.0
서울	265	49.4	271	50.6	536	16.8	63,959	53.0	56,638	47.0	120,596	19.6
부산	96	39.2	149	60.8	245	7.7	25,517	53.4	22,252	46.6	47,769	7.8
대구	111	55.0	91	45.0	202	6.3	20,396	55.8	16,159	44.2	36,556	6.0
인천	93	50.0	93	50.0	186	5.8	18,747	51.7	17,479	48.3	36,227	5.9
광주	79	59.4	54	40.6	133	4.2	10,742	52.3	9,793	47.7	20,534	3.3
대전	48	48.5	51	51.5	99	3.1	10,997	54.2	9,277	45.8	20,274	3.3
울산	51	48.6	54	51.4	105	3.3	8,986	54.8	7,406	45.2	16,391	2.7
경기	275	41.3	391	58.7	666	20.9	69,874	52.0	64,465	48.0	134,339	21.9
강원	57	55.3	46	44.7	103	3.2	9,748	51.9	9,021	48.1	18,769	3.1
충청	123	50.2	122	49.8	245	7.7	22,168	52.8	19,791	47.2	41,959	6.8
전라	136	59.4	93	40.6	229	7.2	24,724	52.0	22,865	48.0	47,589	7.7
경상	260	59.2	179	40.8	439	13.8	39,794	54.4	33,391	45.6	73,185	11.9

Ⅳ. s차년도 종단면 가중치 부여 방법

종단면 분석을 위해 산출되는 s차년도 종단면 가중치는 분석을 위한 대상인 1, 2, …, $s-1$, s차년도에 모두 응답한 학생들만을 대상으로 계산되었다($s=2, 3, \cdots, 6$). 각 차년도의 조사에서 전입/전출, 전학 등으로 이동이 있는 경우에도 해당 학생은 표본으로 계속 유지되어 조사되었으며, $s(=2, \cdots, 6)$차년도의 조사에서 이전 무응답에 대한 원인 등을 파악하기 위한 별도의 추적조사가 실시되지는 않았다. 각 연도별 종단면 가중치 산출방법은 동일함으로 본 논문에서는 3차년도의 예를 들어 s차년도 종단면 가중치의 산출과정을 설명하였다.

3차년도 종단면 가중치는 1, 2, 3차년도에 모두 응답한 학생들(3,017명)을 대상으로 산출되었다. 종단면 가중치 산출 역시 횡단면 가중치 산출과정과 마찬가지로 2단계에 의하여 계산되어진다. 그러나 횡단면 자료와는 달리 종단면 가중치가 부여되는 대상들의

〈표 6〉 χ^2-검정의 예

(단위: 명)

범 주	응 답		무응답		전체	X^2-값	p-값
	빈 도	%	빈 도	%			
고등졸업 이하	945	96.72	32	3.28	977	8.0259	0.0455
2-3년제 대학 졸업	207	94.95	11	5.05	218		
4년제 대학 졸업	689	94.51	40	5.49	729		
대학원 졸업	866	94.13	54	5.87	920		

가중치 부여 시점까지의 응답 구조가 단조 성격을 나타내기 때문에 응답확률 보정을 위하여 로지스틱 회귀분석을 이용한 응답확률 예측값이 사용되었다. 즉 3차년도 의 경우, 각 학생이 2차년도에 응답하였다는 조건하에서 3차년도에 응답할 조건부 확률에 대한 로지스틱 회귀모형을 적합하고 이 결과를 통해 산출되는 응답확률(propensity score)을 가중치 보정을 위해 사용하였다. 이를 위해 3차년도의 응답여부 변수를 응답의 경우 '1' 그리고 무응답의 경우 '0'을 갖는 이항변수로 정의하고 로지스틱 회귀모형을 이용한 응답확률 예측방법을 적용하였다.

로지스틱 회귀모형의 적합에 앞서 주요 설명변수의 응답률 예측에 대한 효율성을 알아보기 위하여 먼저 각 설명변수에 대하여 3차년도 응답자와 무응답자 간의 차이를 비교하는 것이 필요하다. 응답자 그룹과 무응답자 그룹 간의 차이를 20여개의 각 변수별로 분석하기 위해 근사 χ^2-검정을 시행하였다. 각 범주별 빈도수가 적어서 χ^2-검정의 결과를 신뢰할 수 없는 경우 빈도수가 적은 범주를 묶은 후에 검정을 실시하였다. 〈표 6〉은 '학생의 부모님은 학생이 어느 단계까지 학교를 다니기를 바라십니까?'라는 변수에 대한 χ^2-검정의 결과를 제시한 것이다.

그 다음으로 여러 설명변수들의 동시적 예측력을 반영하며 동시에 최대우도 추정량의 유일성을 살피기 위하여 로지스틱 회귀모형의 변수선택 방법을 고려하였다. 변수선택을 위한 방법으로 각 변수들의 추가 또는 제거 후 모형의 설명력에 대한 Wald의 적합도 검정을 바탕으로 한 단계적 방법(stepwise method)을 사용하였다(유의수준=0.15). 〈표 7〉과 〈표 8〉은 변수선택 과정을 통해 선택된 변수들과 로지스틱 회귀분석의 결과를 보여주고 있다. 2, …, 6차년도 각각의 데이터에 대하여 동일한 절차의 분석이 수행되었으며, 차년도별로 약간의 변화가 있기는 하였으나 대체로 〈표 7〉에 제시된 변수들이 주로 선택

〈표 7〉 로지스틱 회귀분석을 이용한 변수선택 결과

변 수	자유도	χ^2-값	p-값
A. 지역(서울, 부산, 대구, …, 충청, 전라, 경상)	11	26.2487	0.0060
B. 학생의 부모님은 학생이 어느 수준까지 교육받기를 원합니까? (고졸 이하, 초대졸, 대졸, 대학원졸)	3	8.5916	0.0352
C. 과목별 반 성적 정도-영어(못한다, 보통, 잘한다)	2	6.9994	0.0302
D. 과목별 반 성적 정도-수학(못한다, 보통, 잘한다)	2	8.0005	0.0183
E. 학원, 과외 등 사교육 수강 여부-국어(없다, 있다)	1	11.4747	0.0007
F. 학원, 과외 등 사교육 수강 여부-영어(없다, 있다)	1	2.4297	0.1191
G. 학원, 과외 등 사교육 수강 여부-수학(없다, 있다)	1	4.8868	0.0271
H. 친한 친구들과는 일주일에 며칠 정도 만납니까? (거의 매일, 2~3일에 한 번, 기타)	2	13.881	0.0010
I. 가출경험이 있습니까?(없다, 있다)	1	7.1363	0.0076
J. 학생은 학생의 삶에 전반적으로 얼마나 만족하고 있습니까? (아니다, 보통, 그렇다)	2	8.9069	0.0116

되었다. 〈표 9〉에는 〈표 8〉의 추정치들을 이용하여 로짓 함수에 의해 얻어진 응답확률 예측값에 대한 요약 통계량이 제시되어 있다. 응답확률의 예측값은 대략 0.515~0.996의 범위 내에 존재하며, 평균과 표준편차는 각각 0.946과 0.044 정도이다.

각 학생에 대한 s차년도 종단면 기본 가중치는 전년도의 종단면 가중치에 이러한 방식으로 계산된 해당 학생의 응답확률 예측값의 역수를 곱하여 계산되었다. 즉,

$$\alpha_{s,hijk}^{l} = w_{s-1,hijk}^{l}\,\hat{p}_{s,hijk}^{(l)\,-1} \tag{5}$$

여기서 $w_{s-1,hijk}^{l}$는 $s-1$차년도 종단면 가중치, $\hat{p}_{s,hijk}^{(l)}$는 로지스틱 회귀분석을 통해 산출된 학생의 응답확률 예측값을 나타낸다.

다음으로 횡단면 가중치 산출과정과 마찬가지로 모집단 변화를 반영하기 위하여 사후층화를 통한 종단면 최종가중치가 산출되었다. 산출된 종단면 최종 가중치는 다음과 같다.

$$w_{s,hijk}^{l} = \alpha_{s,hijk}^{l}\left(\frac{N_{s,p}}{\hat{N}_{s,p}}\right) \tag{6}$$

여기서 $N_{s,p} = \sum_{U} z_{s,p,hijk}$, $\hat{N}_{s,p} = \sum_{S} \alpha_{s,hijk}^{l} z_{s,p,hijk}$, 그리고 $z_{s,p,hijk}$는 h번째 층 내

i번째 학교의 j번째 학급에 속한 k번째 중학생이 s차년도에 사후 층 p에 속하면 1, 그렇지 않은 경우 0의 값을 갖는 지시변수이다.

〈표 8〉 로지스틱 회귀분석의 추정치

변 수	범 주	추정치	표준오차	χ^2-값	p-값	오즈비
Intercept	-	2.2631	0.2144	111.4500	<.0001	
A	1	-0.5492	0.1951	7.9265	0.0049	0.316
	2	-0.1442	0.3096	0.2170	0.6413	0.474
	3	-0.5840	0.2629	4.9346	0.0263	0.305
	4	-0.4255	0.3013	1.9946	0.1579	0.357
	5	-0.3629	0.3574	1.0312	0.3099	0.381
	6	0.8866	0.6694	1.7543	0.1853	1.328
	7	0.5277	0.5575	0.8959	0.3439	0.927
	8	-0.4877	0.1837	7.0463	0.0079	0.336
	9	-0.4253	0.3851	1.2198	0.2694	0.358
	10	1.0869	0.4774	5.1824	0.0228	1.622
	11	-0.1253	0.3103	0.1631	0.6863	0.483
B	1	-0.0464	0.2450	0.0359	0.8497	1.253
	2	0.0285	0.2841	0.0101	0.9201	1.351
	3	0.2902	0.1465	3.9214	0.0477	1.755
C	1	0.1816	0.1309	1.9268	0.1651	1.687
	2	0.1599	0.1247	1.6445	0.1997	1.651
D	1	0.0051	0.1329	0.0015	0.9693	0.745
	2	-0.3040	0.1196	6.4653	0.0110	0.547
E	1	-0.3645	0.1076	11.4747	0.0007	0.482
F	1	-0.2372	0.1522	2.4297	0.1191	0.622
G	1	0.3336	0.1509	4.8868	0.0271	1.949
H	1	0.4256	0.1142	13.8809	0.0002	1.913
	2	-0.2026	0.1495	1.8348	0.1756	1.021
I	1	0.4182	0.1566	7.1363	0.0076	2.308
J	1	-0.4272	0.1533	7.7613	0.0053	0.479
	2	0.1194	0.1198	0.9918	0.3193	0.828

〈표 9〉 응답확률 예측값의 요약 통계량

(단위: 명)

n	최소값	최대값	평 균	표준편차
3,188	0.51556	0.99661	0.94636	0.04406

이러한 절차를 통하여 산출된 s(=2, …, 6)차년도 종단면 가중치의 요약 통계량이 〈표 10〉에 제시되어 있으며, 3차년도 종단면 가중치의 지역별·성별 분포가 〈표 11〉에 제시되어 있다.

〈표 10〉 s차년도 종단면 가중치의 요약 통계량

(단위: 명)

s		n	최소값	최대값	평 균	합 계	표준편차	변동계수
2	전체	3,188	93	488	193	614,192	46	24
	남자	1,594	114	488	204	325,655	50	25
	여자	1,594	93	379	181	288,536	38	21
3	전체	3,017	96	508	203	611,383	51	25
	남자	1,509	118	508	216	325,344	54	25
	여자	1,508	96	402	190	286,039	43	23
4	전체	2,910	100	524	206	598,055	53	26
	남자	1,452	118	524	219	317,401	58	27
	여자	1,458	100	437	192	280,654	45	23
5	전체	2,721	104	647	217	589,925	62	29
	남자	1,370	115	647	227	311,487	68	30
	여자	1,351	104	511	206	278,438	54	26
6	전체	2,459	108	899	240	589,269	78	32
	남자	1,169	124	899	267	311,567	88	33
	여자	1,290	108	553	215	277,702	57	27

〈표 11〉 3차년도 종단면 가중치의 지역별·성별 분포

(단위: 명)

	표본 분포						추정량 분포					
	남 자		여 자		전 체		남 자		여 자		전 체	
	n	%	n	%	n	%	N	%	N	%	N	%
전체	1,490	49.9	1,493	50.1	2,983	100.0	321,266	53.2	282,935	46.8	604,200	100.0
서울	249	50.2	247	49.8	496	16.1	62,842	52.5	56,961	47.5	119,802	19.8
부산	93	39.4	143	60.6	236	7.6	25,144	53.9	21,505	46.1	46,648	7.7
대구	111	57.5	82	42.5	193	6.3	19,828	55.7	15,743	44.3	35,571	5.9
인천	80	47.1	90	52.9	170	5.5	18,083	52.3	16,487	47.7	34,570	5.7
광주	59	53.2	52	46.8	111	3.6	10,578	52.2	9,694	47.8	20,272	3.4
대전	44	47.8	48	52.2	92	3.0	10,873	53.8	9,331	46.2	20,205	3.3
울산	49	48.5	52	51.5	101	3.3	8,975	55.8	7,107	44.2	16,082	2.7
경기	256	41.4	362	58.6	618	20.0	69,500	52.5	62,765	47.5	132,265	21.9
강원	53	55.2	43	44.8	96	3.1	9,647	52.0	8,889	48.0	18,536	3.1
충청	117	50.0	117	50.0	234	7.6	22,145	53.1	19,551	46.9	41,695	6.9
전라	128	59.0	89	41.0	217	7.0	24,587	52.3	22,405	47.7	46,992	7.8
경상	251	59.9	168	40.1	419	13.6	39,064	54.6	32,498	45.4	71,561	11.8

Ⅴ. 토 의

본 논문에서는 한국청소년패널조사 데이터의 분석을 위해 필수적인 횡단면 및 종단면 가중치 산출과정을 소개하였다. 조사 데이터를 이용한 각 조사 연차별 횡단면 통계작성을 위해서는 산출된 횡단면 가중치가 사용되어야 하며, 시계열 분석과 같은 종단면 분석을 위해서는 종단면 가중치가 사용되어야 할 것이다. 종단면 가중치의 경우, 시간이 경

과함에 따라 패널 마모 비율이 높아질 수 있으며 이로 인하여 가중치의 변동이 커지게 된다. 지나치게 크거나 작은 가중치가 산출될 경우, 이를 이용한 통계량의 분산이 커지게 되어 추정량의 통계적 정도가 떨어지는 문제가 발생할 수 있다. 이러한 경우, 로지스틱 분석 시 설명변수의 수나 사후층의 수를 줄이는 방안과 가중치의 상·하한값을 제한하는 calibration 기법을 고려할 수 있다.

참고문헌

김경동 · 이온죽. 1993. ≪사회조사연구방법≫. 서울: 박영사.

김영석. 1999. ≪사회조사방법론≫. 서울: 나남출판.

이경상 · 안선영. 2009. ≪한국청소년패널조사VII : 1-6차년도 조사개요보고서≫.

한국청소년정책연구원. 2009. ≪한국청소년패널조사 중2 패널 1-6차년도 User's Guide≫.

Babbie, Earl. 2001. *The Practice of Social Research*.. ≪사회조사방법론≫ (고성호 외 역, 2002.). 서울: 도서출판 그린.

Groves, R., D. Dillman, J.L. Eltinge, and R.J.A. Little. 2002. *Survey Nonresponse*. New York: Wiley.

Sarndal, C.E. and S. Lundstorm. 2006. *Estimation in Survey with Nonresponse*. New York: Wiley.

<접수 2011/10/26, 수정 011/11/22, 게재확정 2011/11/23>