세미나 23-02

다문화 청소년 패널조사(MAPS)

데이터 설명회 및 방법론 특강

2023. 8. 24.(목) 10:00~12:20

장 소 대한상공회의소 의원회의실

주최 한국청소년정책연구원

다문화 청소년 패널조사(MAPS)

데이터 설명회 및 방법론 특강

프로그램

시간	내용		비고
		사회 이정민 (한국청소년정책연구원 부연구위원)	
10:00~ 10:10	개 회	| 개회사 | 김현철 (한국청소년정책연구원 원장)	
10:10~ 10:40	데이터 설명회	| 발 표 | 다문화 청소년 종단연구(2기)의 구조와 활용방안 발표자 신동훈 (한국청소년정책연구원 부연구위원)	유튜브 동시 송출
10:40~ 12:20	데이터 활용방법론 특강	| 발표 1 | 패널자료 연구방법론의 이론 발표자 김현식 (경희대학교 교수)	
		| 발표 2 | R과 STATA를 활용한 다문화 청소년 패널조사 분석 발표자 김현식 (경희대학교 교수)	

차례

다문화 청소년 패널조사(MAPS)

데이터 설명회 및 방법론 특강

데이터 설명회

- **신동훈** (한국청소년정책연구원 부연구위원)

「다문화 청소년 패널조사(MAPS)
데이터 설명회 및 방법론 특강」
2023. 08. 24.

다문화 청소년 패널조사: 자료 설명과 향후 계획

한국청소년정책연구원 신동훈 부연구위원

이 발표자료는 2023년 5월에 개최된 제15회 KOSSDA 데이터페어 '변화를 담는 패널데이터'에서 발표된 신동훈의 '다문화 청소년 패널조사: 자료 소개와 데이터 설명'의 내용을 수정보완한 것임.

1

목 차

2

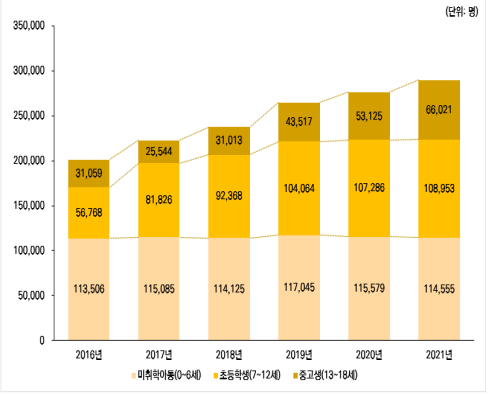
조사개요

3

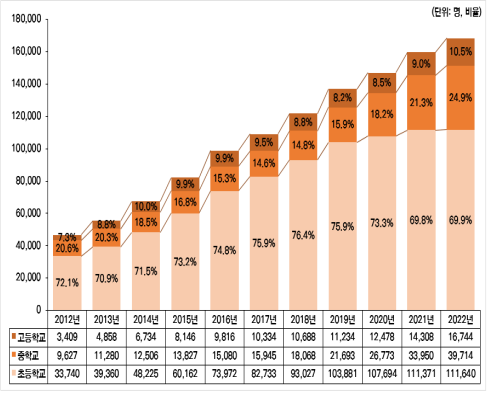
Ⅰ. 조사개요

1. 조사배경

다문화 청소년의 증가: 외국인주민 자녀 / 다문화학생



출처: 행정안전부(2017~2022). (각년도) 지방자치단체 외국인주민 현황



	2012년	2013년	2014년	2015년	2016년	2017년	2018년	2019년	2020년	2021년	2022년
고등학교	3,409	4,858	6,734	8,146	9,816	10,334	10,688	11,234	12,478	14,308	16,744
중학교	9,627	11,280	12,506	13,827	15,080	15,945	18,068	21,693	26,773	33,950	39,714
초등학교	33,740	39,360	48,225	60,162	73,972	82,733	93,027	103,881	107,694	111,371	111,640

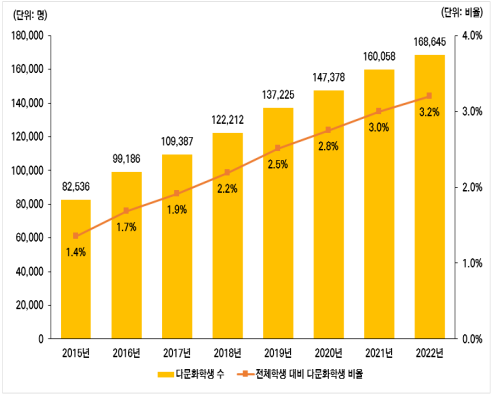
출처: 교육부 보도자료(2020.08.27., 2022.08.31.). (각년도) 교육기본통계 결과 발표

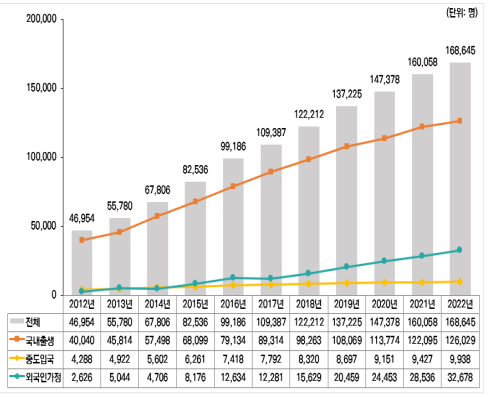
4

Ⅰ. 조사개요

1. 조사배경

다문화학생이 차지하는 비율의 증가 및 다문화학생 유형의 다양화





	2012년	2013년	2014년	2015년	2016년	2017년	2018년	2019년	2020년	2021년	2022년
전체	46,954	55,780	67,806	82,536	99,186	109,387	122,212	137,225	147,378	160,058	168,645
국내출생	40,040	45,814	57,498	68,099	79,134	89,314	98,263	108,069	113,774	122,095	126,029
중도입국	4,288	4,922	5,602	6,261	7,418	7,792	8,320	8,697	9,151	9,427	9,938
외국인가정	2,626	5,044	4,706	8,176	12,634	12,281	15,629	20,459	24,453	28,536	32,678

출처: 교육부 보도자료(2020.08.27., 2022.08.31.), (각년도) 교육기본통계 결과 발표

Ⅰ. 조사개요

2. 조사 목적 및 특징

목적	다문화 청소년의 발달 특성을 객관적이고 체계적으로 이해할 수 있는 기초자료 구축 → 실증적 분석을 바탕으로 근거기반 정책 추진에 기여
특징	다문화 청소년[1]의 다방면에 걸친 발달 궤적[2] 을 관찰하기 용이 → 시간적 선후에 근거한 요인간 인과관계 분석이 가능
비교	1) 한국아동청소년패널(한국청소년정책연구원): 청소년 대상 종단조사 2) 다문화가족실태조사(한국여성정책연구원): 다문화가족 대상 3년 주기 횡단조사

Ⅰ. 조사개요

3. 조사대상

1기 코호트

- 다문화 청소년과 그들의 어머니
  - ✓ 2011년 기준 초등학교 4학년 재학 다문화 청소년 대상
- 다문화 청소년의 의미
  - ✓ 국제결혼가정자녀, 중도입국청소년, 외국인자녀 등을 모두 포함하는 개념으로 국제결혼가정자녀에 한정하고자 한 것은 아니나, 모집단 분포에서 국제결혼가정자녀가 대다수를 차지하기에 결과적으로 본 조사에 국제결혼가정자녀가 주로 포함됨

2기 코호트

- 다문화청소년과 그들의 어머니
  - 2019년 기준 초등학교 4학년 재학 다문화 청소년
- 다문화 청소년의 의미
  - ✓ (교육부 다문화학생 정의에 근거) 국제결혼가정자녀, 중도입국청소년, 외국인자녀 등을 모두 포함하는 개념.
  - ✓ 국제결혼가정자녀: 한국인과 외국인 배우자가 결혼한 가정 자녀 중 국내에서 출생한 자녀
  - ✓ 중도입국청소년: 한국인과 외국인 배우자가 결혼한 가정 자녀 중 외국에서 태어나 일정 기간 성장한 이후 국내에 입국한 자녀
  - ✓ 외국인자녀: 외국인 사이에서 출생한 자녀

7

조사설계 및 관리

8

Ⅱ. 조사설계 및 관리

1. 조사설계 및 모형

- 모집단
  - ✓ 2011년(2019년: 2기) 기준 국내 거주 다문화 청소년 중 초등학교 4학년 학생과 그들의 학부모(어머니)
    - ※ 1기 코호트: 전국 2,537개 초등학교 4학년에 재학 중인 다문화가정 학생 4,452명(2011년 또는 2010년 기준 현황)
    - ※ 2기 코호트: 전국 4,805개 초등학교 4학년에 재학 중인 다문화가정 학생 17,134명(2018년 기준 현황)
- 기본 조사단위
  - ✓ 다문화 청소년 중 초등학교 4학년 학생과 그들의 학부모(어머니)
  - ✓ 표본추출단위는 다문화 청소년 중 초등학교 4학년 학생이 재학 중인 초등학교로 정의
    - ※ 국제결혼가정자녀: 한국인과 외국인 배우자가 결혼한 가정 자녀 중 국내에서 출생한 자녀
    - ※ 중도입국청소년: 한국인과 외국인 배우자가 결혼한 가정 자녀 중 외국에서 태어나 일정 기간 성장한 이후 국내에 입국한 자녀
    - ※ 외국인자녀: 외국인 사이에서 출생한 자녀 (교육부, 한국교육개발원(2018)의 정의 인용)
- 표본추출틀
  - ✓ 2011년 4월 기준(2018년 4월 기준: 2기) 각 지방 교육청의 초등학교 4학년 다문화 청소년이 재학 중인 학교 리스트

9

Ⅱ. 조사설계 및 관리

1. 조사설계 및 모형

- 표본추출 단위: 학교
  - ✓ 추출된 학교에서는 조사 대상에 해당하는 학생 전수조사를 원칙으로 함
- 표본추출 방법
  - ✓ 1단계: 표본학교 추출은 (각 층별로 랜덤하게 추출하는) 층화임의추출법 적용
  - ✓ 2단계: (층 내에서는 다문화가정 학생 수가 많은 학교의 추출율을 크게 하는) 확률비례추출법 적용
- 표본의 크기는 데이터의 신뢰성과 예산을 고려하여 청소년 및 학부모 각 1,600명(2기: 2,100명)으로 결정함
- 구축표본
  - ✓ 1기 코호트: 청소년 1,635명, 학부모 1,625명
  - ✓ 2기 코호트(2차 조사 때 추가 구축한 중도입국청소년 26가구 포함): 청소년 2,271명, 학부모 2,249명
    - ※ 가구기준: 국제결혼가정(1,719가구), 중도입국(175가구), 외국인가정(355가구)

10

Ⅱ. 조사설계 및 관리

2. 조사방법

1기 코호트

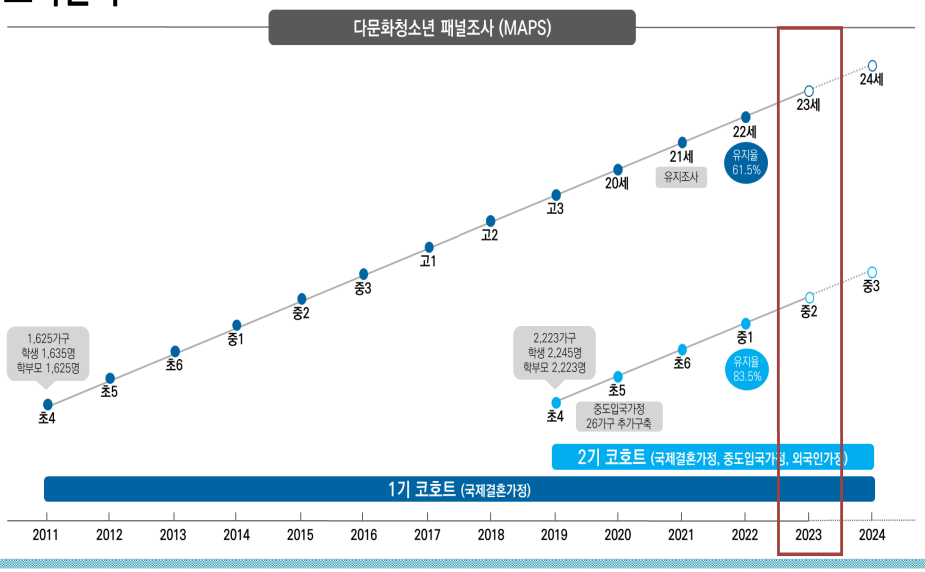
- 전문 조사원에 의한 면접조사 방식
- 조사도구의 변화
  - ✓ CAPI(Computer Assisted Personal Interview): 1~8차 → TAPI(Tablet Assisted Personal Interview): 9~10, 12차 → 전화유지조사(온라인조사 포함): 11, 13차
- 설문지 언어
  - ✓ 학부모(어머니): 한국어 및 총 9개 외국어(영어, 중국어, 몽골어, 필리핀어, 러시아어, 베트남어, 태국어, 일본어, 아랍어 등)로 번역된 설문지 중 직접 언어 선택
  - ✓ 다문화 청소년: 한국어 설문지
- 10차 조사 이후 학부모(어머니) 조사 대폭 축소

2기 코호트

- 전문 조사원에 의한 면접조사 방식
- 조사도구: TAPI(Tablet Assisted Personal Interview)
- 설문지 언어
  - ✓ 1차 조사 시 한국어 및 총 9개 외국어(영어, 중국어, 몽골어, 필리핀어, 러시아어, 베트남어, 태국어, 일본어, 아랍어 등)로 번역된 설문지 중 직접 언어 선택
  - ✓ 2차 조사부터는 번역본 활용도를 고려하여 외국인 학부모 8개 외국어(아랍어 제외), 학생용은 5개 외국어(몽골어, 필리핀어, 태국어 제외)로 번역

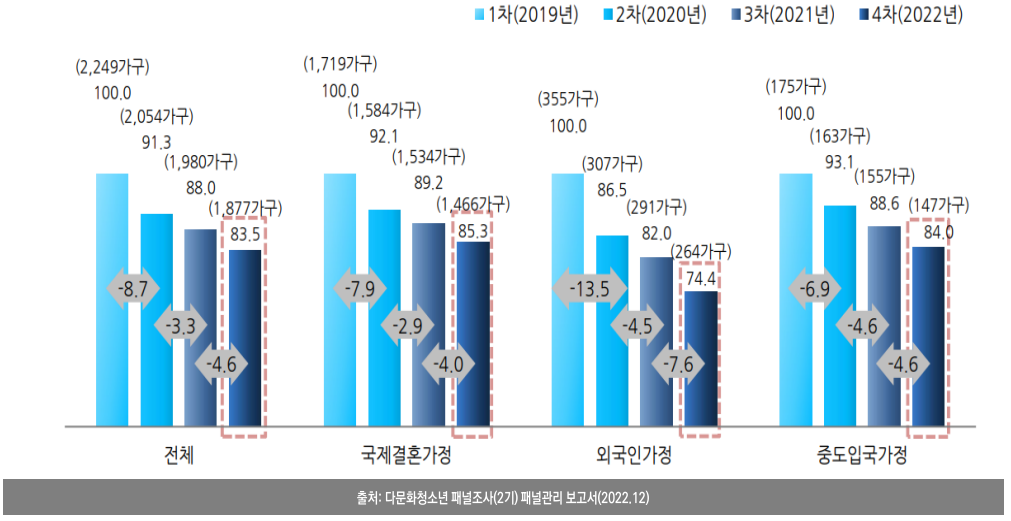
Ⅱ. 조사설계 및 관리

3. 조사관리



Ⅱ. 조사설계 및 관리

3. 조사관리



출처: 다문화청소년 패널조사(2기) 패널관리 보고서(2022.12)

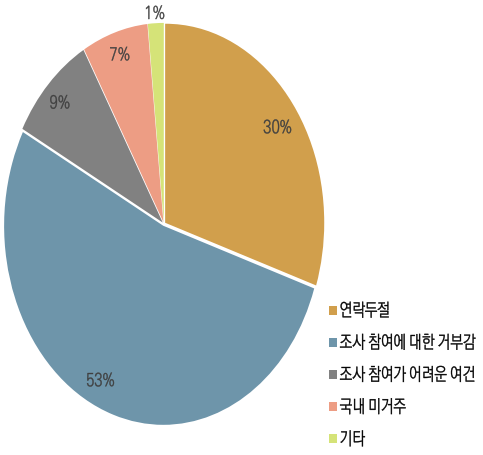
Ⅱ. 조사설계 및 관리

3. 조사관리

- 2019년: 2,223가구 패널 구축
  - ✓ 학생 2,245명, 학부모 2,223명 → 가구 내 조사 대상 학생이 2명인 가구가 22가구
- 2020년: 총 2,028가구(학생 기준) 조사 참여 → 패널 유지율 91.2% (추가 구축 포함 패널 유지율 91.3%)
  - ✓ 어머니: 2,021명(대체양육자 제외), 학생 : 2,047명
  - ✓ 2020년 중도입국가구 26가구 추가 구축
- 2021년: 총 1,980가구(학생 기준) 조사 참여 → 패널 유지율 88.0%
  - ✓ 어머니: 1,970명(대체양육자 제외), 학생: 1,999명
- 2022년: 총 1,877가구(학생 기준) 조사 참여 → 패널 유지율 83.5%
  - ✓ 어머니: 1,856명(대체양육자 제외), 학생: 1,895명

Ⅱ. 조사설계 및 관리

3. 조사관리



- **패널조사 이탈 사유**
  - ✓ 연락두절 : 30.1%
  - ✓ 조사 참여에 대한 거부감 : 52.7%

    ※ 바빠서, 코로나로 인해 거절, 이사 후 주소지 안 알려주고 조사 거절 등
  - ✓ 조사 참여가 어려운 여건 : 8.9%

    ※ 이혼 후 조사 참여 거절, 학생 거부, 부모 중 일부 부재, 학생 건강 등
  - ✓ 국내 미거주 : 6.7%

    ※ 유학, 이민 등
  - ✓ 기타 : 1.6%

15

조사 내용

16

Ⅲ. 조사 내용

1. 청소년(2기)

구분		조사항목	출처	2019(1차년도)		2020(2차년도)		2021(3차년도)		2022(4차년도)	
대영역	소영역			국내	외국	국내	외국	국내	외국	국내	외국
문화적응 및 이중문화	언어능력	부모의 외국인 여부	MAPS 1기	○	○	○	○	○	○	○	○
		외국 출신 부모의 한국어 실력	CILS(Portes & Rumbaut, 2008) 문항 및 MAPS 1기 문항 수정	○	○	○	○	○	○	○	○
		부모와의 의사소통 시 사용 언어	CILS((Portes & Rumbaut, 2008) 문항 및 MAPS 1기 문항 수정	○	○	○	○	○	○	○	○
		외국 출신 부모 나라의 언어 교육	연구진 작성	○	○	○	○	○	○	○	○
		외국 출신 부모 나라 언어 실력	MAPS 1기 문항 수정	○	○	○	○	○	○	○	○
		자신의 한국어 실력	CILS(Portes & Rumbaut, 2008) 문항 수정 및 MAPS 1기 문항	-	○	-	○	-	○	-	○
		입국 후, 학교 입학 전 한국어 배운 경험 및 장소	연구진 작성	-	○	-	-	-	-	-	-
	이중문화 경험	국적에 대한 인식	CILS(Portes & Rumbaut. 2008) 문항 수정 및 MAPS 1기 문항 수정	○	○	○	○	○	○	○	○
		외국 출신 부모 나라 방문여부 및 횟수	MAPS 1기 문항 수정	○	-	○	-	○	-	○	-
		문화적응스트레스	Hovey & King(1996) 번안 및 수정하여 사용한 노충래(2000)를 수정한 홍진주(2004)를 수정하여 사용한 MAPS 1기 수정	○	○	○	○	-	○	-	○
		국가정체성	성한기(2001) 발췌 및 수정	○	○	○	○	○	○	○	○
		이중문화수용태도	노충래, 홍진주(2006) 발췌 및 수정	○	○	○	○	○	○	○	○
	지원정책	특별지원 필요성 및 방식	MAPS 1기 수정	-	-	-	-	-	-	○	○
		원하는 지원내용	MAPS 1기 수정	-	-	-	-	-	-	○	○

Ⅲ. 조사 내용

1. 청소년(2기)…계속

구분		조사항목	출처	2019(1차년도)		2020(2차년도)		2021(3차년도)		2022(4차년도)	
대영역	소영역			국내	외국	국내	외국	국내	외국	국내	외국
개인요인	심리사회적응, 행동 및 건강	건강평가	청소년종합실태조사(백혜정, 임희진, 김현철, 유성렬, 2017) 수정	○	○	○	○	○	○	○	○
		스트레스	청소년종합실태조사(백혜정 외, 2017) 수정	○	○	○	○	○	○	-	-
		정서문제(공격성, 사회적 위축, 우울)	한국아동·청소년패널조사 2018(하형석 외, 2018) 일부 문항 발췌	-	-	-	-	-	-	○	○
		자존감	청소년종합실태조사(백혜정 외, 2017) 수정	○	○	○	○	○	○	○	○
		사회적 역량	청소년종합실태조사(백혜정 외, 2017) 수정	○	○	-	-	-	-	-	-
		삶의 만족도	김신영 외(2006)을 인용한 한국아동·청소년패널 2010(김지경 외, 2010)	○	○	○	○	○	○	○	○
		현재 걱정거리	유한구 외(2016)에서 발췌 및 수정 한 MAPS 1기 수정	○	○	○	○	○	○	○	○
		차별 경험 피해여부 및 가해자, 차별경험 대처	국내 난민아동 한국사회 적응 실태조사(노충래, 고주애, 윤수경, 2018) 수정	○	○	○	○	○	○	○	○
		한국사회 다문화 수용성에 대한 인식	연구진 작성	○	○	○	○	○	○	○	○
		비행	이경상 외(2003)를 수정·보완한 이경상 외(2011) 및 한국아동·청소년패널 2018(하형석 외, 2018)을 참고하여 수정 및 보완하여 사용	-	-	-	-	-	-	○	○
		사이버비행	청소년 사이버폭력의 유형분석 및 대응방안 연구(이승현, 강지현, 이원상, 2015) 사이버비행 가해경험을 수정 사용한 한국아동·청소년패널조사 2018(하형석 외, 2018) 사이버비행 중 사이버비방 발췌	-	-	-	-	-	-	○	○
	인지	학교성적	한국아동·청소년패널조사 2018(하형석 외, 2018)	○	○	○	○	○	○	○	○
		성적에 대한 만족도	MAPS 1기	○	○	○	○	○	○	○	○
	진로	희망교육수준	MAPS 1기	○	○	○	○	○	○	○	○
		희망직업	MAPS 1기	○	○	○	○	○	○	○	○
		학교 마친 후 일하고 싶은 나라	연구진 구성	○	○	○	○	○	○	○	○
		미래, 진로, 진학 관련 대화(상담)하는 사람	한국아동·청소년패널조사 2018(하형석 외, 2018)	○	○	○	○	○	○	○	○
		진로태도 결정성	이기학, 한종철(1997)의 진로태도 중 진로태도 결정성 문항 수정·보완하여 사용	-	-	-	-	-	-	○	○

Ⅲ. 조사 내용

1. 청소년(2기)…계속

구분		조사항목	출처	2019(1차년도)		2020(2차년도)		2021(3차년도)		2022(4차년도)	
대영역	소영역			국내	외국	국내	외국	국내	외국	국내	외국
환경요인	부모	부모님의 지지(교육적 지원 및 기대)	김순규(2001) 수정	○	○	○	○	○	○	○	○
		부모의 양육태도	허묘연(2000) 수정 발췌한 한국아동·청소년패널조사 2010(김지경 외, 2010, 이경상 외, 2011)	○	○	○	○	○	○	○	○
		부모님과 활동 및 대화	청소년종합실태조사(백혜정 외, 2017) 수정	○	○	○	○	○	○	○	○
		하루 부모님과의 대화 시간	청소년종합실태조사(백혜정 외, 2017) 수정	○	○	○	○	○	○	○	○
		방과 후 보호자 부재	청소년종합실태조사(백혜정 외, 2017)	○	○	○	○	○	○	○	○
		가정형편	연구진 작성	○	○	–	–	–	–	–	–
	친구	친구의 지지	한미현(1996) 발췌	○	○	○	○	○	○	○	○
		집단괴롭힘 피해경험	이해경, 김혜원(2001) 수정	○	○	○	○	○	○	○	○
		친한 친구 수	MAPS 1기	○	○	○	○	○	○	○	○
		친한 친구 국적	MAPS 1기 및 연구진 작성	–	○	–	○	–	○	–	○
	학교생활	학교생활	청소년종합실태조사(백혜정 외, 2017)	○	○	○	○	○	○	○	○
		학교생활 전체의 어려운 점	연구진 작성	○	○	○	○	○	○	○	○
		학교생활 어려운 점 (친구)	MAPS 1기 수정	○	○	○	○	○	○	○	○
		학교공부의 어려운 점	MAPS 1기 수정	○	○	○	○	○	○	○	○
		교사의 지지	한미현(1996) 발췌 및 수정	○	○	○	○	○	○	○	○
		학교내 도움 받을 수 있는 사람	MAPS 1기 수정	○	○	○	○	○	○	○	○
	지역사회 지지망	학교밖에서 도움 받을 수 있는 사람	MAPS 1기 수정	○	○	○	○	○	○	○	○
	학업중단 위험요인	학업중단 위험요인 생활 및 태도	이자영 외(2010) 문항 수정. 보완하여 사용	–	–	–	–	–	–	○	○
		학업중단 위험요인 친구 및 선후배 영향	김순규(2001)의 문항 수정. 보완하여 사용	–	–	–	–	–	–	○	○

Ⅲ. 조사 내용

1. 청소년(2기)…계속

구분		조사항목	출처	2019(1차년도)		2020(2차년도)		2021(3차년도)		2022(4차년도)	
대영역	소영역			국내	외국	국내	외국	국내	외국	국내	외국
중도입국	학교관련	한국 오기 전 학교 경험	연구진 작성	–	○	–	–	–	–	–	–
		한국 입학(편입) 학년	연구진 작성	–	○	–	–	–	–	–	–
		한국 입국 후 초등학교 입학 전 생활	연구진 작성	–	○	–	–	–	–	–	–
	미래 생활	국적취득 희망	연구진 작성	–	○	–	○	–	○	–	○
		한국 거주 희망	연구진 작성	–	○	–	○	–	○	–	○
기타	방과후생활	방과 후 활동	청소년종합실태조사(백혜정 외, 2017)	○	○	–	–	–	–	–	–
	청소년 활동	문화생활, 스포츠 관람, 레저시설 이용 등	사회조사표(통계청, 2017) 발췌 및 수정	○	○	–	–	–	–	–	–
		청소년 활동 참여 경험	한국아동·청소년패널조사 2018(하형석 외, 2018) 수정·보완하여 사용	–	–	–	–	–	–	○	○

Ⅲ. 조사 내용

2. 청소년(2기) 신규문항

구분		문항수	2023년 (중2)	2024년 (중3)	2025년 (고1)	2026년 (고2)	2027년 (고3)	조사 주기	유사영역 존재여부
대영역	세부영역								
개인요인	진로준비역량	12문항	●		●		●	2년 (홀수해)	○ (진로, 진로결정성)
	상급학교 진학 준비도	7문항	●	●	●	●	●	상시	X
학습관련 심리정서	그릿	8문항		●		●		2년 (짝수해)	X
환경요인	방과후 생활	4문항		●		●		2년 (짝수해)	△
환경요인	인권	4문항	●	●	●	●	●	상시	X
개인요인	사회적 고립(외로움)	4문항	●	●	●	●	●	상시	○ (심리사회적응,삶의 만족도)
개인요인	학교생활	1문항	●	●	●	●	●	상시	비행관련 문항에 온라인 도박 추가

Ⅲ. 조사 내용

3. 학부모(2기)

구분		조사항목	출처	2019 (1차년도)				2020 (2차년도)			2021 (3차년도)			2022 (4차년도)		
				외국인학부모		한국인학부모		외국인	한국인학부모		외국인	한국인학부모		외국인	한국인학부모	
대영역	소영역			국내	외국	국내	외국	학부모	국내	외국	학부모	국내	외국	학부모	국내	외국
배경변인	가족특성	배우자와의 결혼 상태 및 관계 만족도	MAPS 1기 및 국내 난민아동 한국사회 적응 실태조사(노충래 외, 2018) 수정	○	○	○	○	○	○	○	○	○	○	○	○	○
		가정의 월평균 소득	MAPS 1기 수정	○	○	○	○	○	○	○	○	○	○	○	○	○
		가정의 주요 소득원	MAPS 1기	○	○	○	○	○	○	○	○	○	○	○	○	○
		가정의 경제적 수준(가정형편)	MAPS 1기	○	○	○	○	○	○	○	○	○	○	○	○	○
개인요인	심리사회적 응 및 건강	차별 경험 피해여부 및 장소, 차별 대처	국내 난민아동 한국사회 적응 실태조사(노충래 외, 2018) 수정	○	○	–	–	–	–	–	–	–	–	–	–	–
		차별 경험 피해여부 및 대상별 피해 경험, 차별 대처 및 차별 사유	김지혜 외(2019) 및 한준 외(2019) 발췌 및 수정, 연구진 작성	–	–	–	–	○	–	–	○	–	–	○	–	–
		자아존중감	Rosenberg(1965) 발췌	○	○	○	○	○	○	○	○	○	○	○	○	○
		전반적 건강 상태	MAPS 1기	○	○	○	○	○	○	○	○	○	○	○	○	○
		걱정거리 의논할 수 있는 사람	Kim, U. (1988) 수정 및 보완 한 MAPS 1기	○	○	○	○	○	○	○	○	○	○	○	○	○
문화특성	언어능력	모국어	MAPS 1기	○	○	–	–	○	–	–	○	–	–	○	–	–
		한국어 학습경험	MAPS 1기	○	○	–	–	○	–	–	○	–	–	○	–	–
		한국어 수준	CILS(Portes & Rumbaut, 2008) 문항 수정	○	○	–	–	○	–	–	○	–	–	○	–	–
	문화적응	주로 어울리는 사람	Kim, U. (1988) 수정 및 보완 한 MAPS 1기	○	○	–	–	○	–	–	○	–	–	○	–	–
		한국생활이 모국문화와 달라서 어려운 점	Kim, U. (1988) 수정 및 보완 한 MAPS 1기	○	○	–	–	○	–	–	○	–	–	○	–	–
		문화적응 스트레스	이승종(1995) 번안, 이소래(1997)가 수정한 것을 수정 후 사용한 MAPS 1기 수정	○	○	–	–	○	–	–	○	–	–	○	–	–
		문화적응 유형	Barry(2001) 발췌 및 수정 한 MAPS 1기 수정	○	○	–	–	○	–	–	○	–	–	○	–	–

Ⅲ. 조사 내용

3. 학부모(2기)…계속

구분		조사항목	출처	2019(1차년도)				2020(2차년도)			2021(3차년도)			2022(4차년도)		
대영역	소영역			외국인학부모		한국인학부모		외국인학부모	한국인학부모		외국인학부모	한국인학부모		외국인학부모	한국인학부모	
				국내	외국	국내	외국		국내	외국		국내	외국		국내	외국
자녀 양육	자녀 교육 및 양육	자녀와 대화 시 사용 언어	MAPS 1기	○	○	–	○	○	–	○	○	–	○	○	–	○
		자녀와의 의사소통 정도	MAPS 1기	○	○	–	○	○	–	○	○	–	○	○	–	○
		자녀의 모국(본국)어 능력	연구진 작성	○	○	–	○	○	–	–	○	–	–	○	–	–
		초등학교 입학 전 주 양육자	MAPS 1기 수정	○	○	○	○	–	–	–	–	–	–	–	–	–
		한국오기전자녀부모님과떨어져산기간및그기간중인함께한사람	연구진 작성	–	○	–	○	–	–	–	–	–	–	–	–	–
		입국후초등학교입학기간및소요이유	연구진 작성	–	○	–	○	–	–	–	–	–	–	–	–	–
		초등학교 입학 시기 및 입학이 늦어진 사유	연구진 작성	○	–	○	–	–	–	–	–	–	–	–	–	–
		자녀교육 관련 주 의사결정자	MAPS 1기 수정	○	○	○	○	○	○	○	○	○	○	○	○	○
		자녀 학교교육에 대한 도움 정도	MAPS 1기	○	○	○	○	○	○	○	○	○	○	○	○	○
		자녀 학교문제, 숙제지도 하는 사람	MAPS 1기	○	○	○	○	○	○	○	○	○	○	○	○	○
		자녀의 성적향상을 위한 노력	MAPS 1기 수정	○	○	○	○	○	○	○	○	○	○	○	○	○
		자녀의 학교생활을 위해 하는 활동	MAPS 1기 수정	○	○	○	○	○	○	○	○	○	○	○	○	○
		선호하는 자녀 교육 방식	MAPS 1기 수정	○	○	○	○	○	○	○	○	○	○	○	○	○
		방과후 교육(사교육 경험)	한국아동·청소년패널조사 2010(김지경 외, 2010) 수정	○	○	○	○	○	○	○	○	○	○	○	○	○
	양육 관련 특성	자녀 양육 시 어려운 점	국내 난민아동 한국사회 적응 실태조사(노충래 외, 2018)	○	○	○	○	○	○	○	○	○	○	○	○	○
		부모효능감	Gibaud-Wallston&Wandersman(1978)의문항을 변인및수정한산속재(1997)을수정한MAPS 1기수정	○	○	○	○	○	○	○	○	○	○	○	○	○
	자녀 진학/진로 및 향후 희망 거주	희망하는 자녀 교육 수준	MAPS 1기 수정	○	○	○	○	○	○	○	○	○	○	○	○	○
		희망하는 자녀 거주 방식	연구진 작성	○	○	○	○	○	○	○	○	–	–	○	–	–
		자녀가 나중에 갖기 원하는 직업	MAPS 1기	○	○	○	○	○	○	○	○	○	○	○	○	○
다문화 지원 정책	지원 정책 태도	지원프로그램 제공정보 선호도	MAPS 1기	○	○	○	○	○	○	○	○	○	○	○	○	○
		자녀 다문화가족정책 지원 경험 및 도움 정도	김영혜, 김다경(2016)	○	○	○	○	○	○	○	○	○	○	○	○	○
		선호하는 지원 방식	MAPS 1기 수정	○	○	○	○	○	○	○	○	○	○	○	○	○

Ⅲ. 조사 내용

4. 학부모(2기) 신규문항

구분		문항수	2023년 (중2)	2024년 (중3)	2025년 (고1)	2026년 (고2)	2027년 (고3)	조사 주기	유사영역 존재여부
대영역	세부영역								
자녀 양육	자녀 진로준비를 위한 학부모 활동	5문항	●	●	●	●	●	상시	X
	자유학년제 (또는 자유학기제)	1문항	●					일회성	X
	상급학교 진학 준비도	8문항	●	●	●	●	●	상시	X

데이터 활용

25

Ⅳ. 데이터 활용

1. 발간 논문 실적

구분		활용실적						
		2017	2018	2019	2020	2021	20221	전체
학술지	전문 학술지	9	13	49	86	115	76	348
학위	석사학위	-	-	7	16	17	17	57
	박사학위	-	-	5	11	6	10	32
학술대회 발표	MAPS	32	-	-	-	-	-	32
	기타	1	1	5	2	4	6	19
기타(타 기관 연구보고서 외)		-	-	1	1	1	-	3
계		42	14	67	116	143	90	491

출처: 양계민 외(2022: 32) 표 II-1 수정

26

Ⅳ. 데이터 활용

2. 발간 논문 소개

순위	주요어	빈도	순위	주요어	빈도
1	다문화청소년	1,466	11	자녀	248
2	자아존중감	512	12	청소년	229
3	문화적응스트레스	488	13	매개효과	215
4	어머니	390	14	다문화가정	207
5	학교생활적응	380	15	삶의 만족도	180
6	우울	324	16	이중문화수용태도	178
7	부모	278	17	진로장벽	164
8	사회적위축	274	18	자아탄력성	160
9	성취동기	271	19	진로결정	158
10	집단	266	20	중학교	155

* 주: 빈도는 20위까지만 제시함

출처: 양계민 외(2022: 34) 그림 II-3 & 표 II-2

Ⅳ. 데이터 활용

3. 분석 유의점

- 한가구에 청소년이 2명인 사례 존재
  - ✓ 쌍둥이 혹은 가구 내 같은 학년에 재학 중인 청소년이 2명인 사례가 있어 청소년과 학부모 수에 차이가 있음
- 10차 조사 이후 1기 코호트 조사 방식
  - ✓ 유지조사(11차:2021년) – 본조사(12차: 2022년) – 유지조사(13차: 2023년) – ~~본조사 계획(14차: 2024년)~~
- 중도입국청소년 및 외국인 자녀
  - ✓ 대상의 특성상 초기 구축된 모집단의 특성이 상황에 따라 달라질 수 있음

Ⅳ. 데이터 활용

3. 분석 유의점

• 본 자료를 활용하여 학술논문을 발표할 경우 결과 해석이 특정 국가 및 다문화에 대한 고정관념과 편견을 강화시키는 데 부정적 영향을 미칠 가능성을 숙고함.

□ 예 □ 아니오

• 본 자료를 분석한 결과를 해석할 경우 다문화 청소년과 비다문화 청소년의 결과를 집단 대 집단으로 단순하게 비교하여 발표하는 것에 주의함.

□ 예 □ 아니오

• 본 자료를 분석한 결과를 발표할 때는 그 결과가 다문화적 특성에 기인한 것인지 사회경제적 배경에서 기인한 것인지 등을 면밀히 분석한 후 제시함.

□ 예 □ 아니오

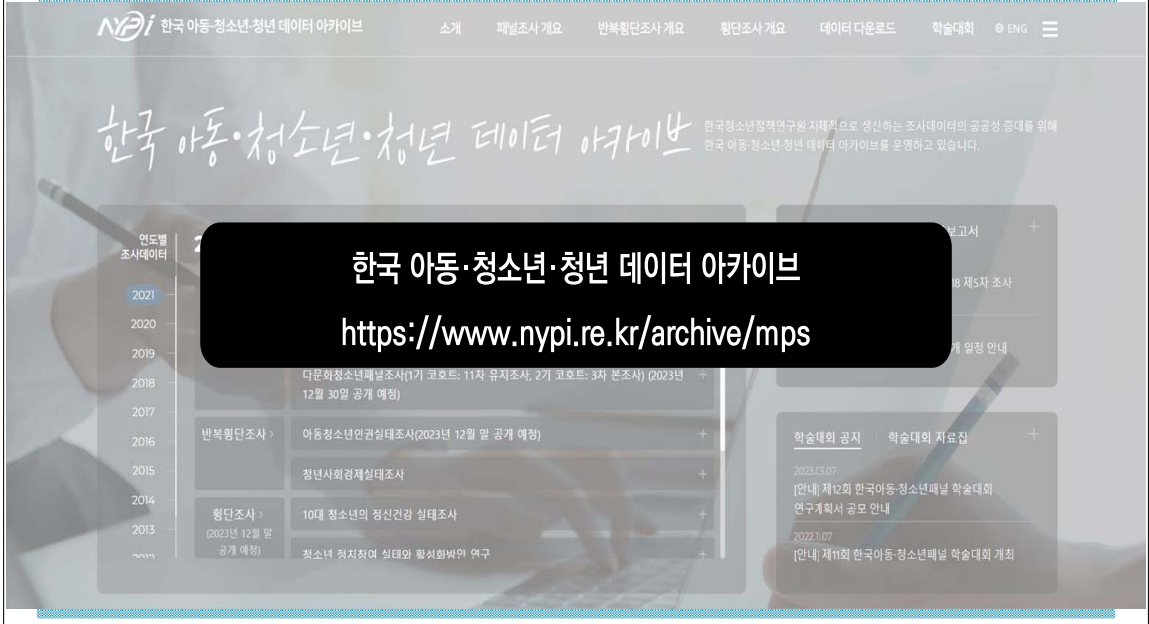
• 본 자료를 분석한 논문결과 보고 시 인종차별적인 표현을 사용하지 않도록 유의함.

□ 예 □ 아니오

• 본 자료를 분석한 연구결과가 다문화청소년과 다문화 가정에 미칠 수 있는 영향에 대해 숙고함.

□ 예 □ 아니오

Ⅳ. 데이터 활용



Ⅳ. 데이터 활용

4. 향후 계획(데이터 신청)

데이터 신청

maps@nypi.re.kr

2023년 연말 – 2024년 초 공개방식 변경

한국 아동·청소년·청년 데이터 아카이브

직접 다운로드

31

감사합니다

다문화 청소년 패널조사(MAPS)

데이터 설명회 및 방법론 특강

데이터 활용방법론 특강

- **김현식** (경희대학교 교수)

고정효과모형, 확률효과모형과 다층모형

다문화청소년패널조사 데이터 설명회
2023.08.24

김현식
sochyunsik@khu.ac.kr
경희대학교 사회학과

1

목차

2

1: 자료 소개

- 연구 가설
- 종단자료 만들기

3

연구 가설 1

- 외국출신여성의 한국어 능력이 자녀의 자기보고 학교성적에 미치는 영향
- 연구 자료
  - https://www.nypi.re.kr/archive/board?menuId=MENU00221&siteId=null
  - 다문화청소년패널조사 1기 1-10차 자료 중 1-9차 자료를 사용
  - 한국청소년정책연구원에 자료를 신청한 후 승인 절차를 거쳐 자료를 받아 특정 폴더에 저장한다
  - 한국의 많은 자료들이 SPSS 파일로 만들어져 있기 때문에 여기서는 SPSS 파일을 받았다고 가정하고 논의를 진행한다
  - 패널의 유저가이드, 통합설문지, 통합코드북은 웹사이트에서 모두 다운받을 수 있다

4

연구 가설 2

- 설명 변수

4 귀하는 현재 한국어를 얼마나 잘 할 수 있습니까?

구분	전혀 못함	못하는 편	잘하는 편	매우 잘함
말하기	1	2	3	4
쓰기	1	2	3	4
읽기	1	2	3	4
듣기	1	2	3	4

- 종속 변수

29 다음은 학생의 성적에 관한 질문입니다. 각 과목에 대해 자신의 성적이 해당되는 번호를 솔직하게 골라주세요.

번호	과 목	매우 못하는 편이다	못하는 편이다	보통이다	잘하는 편이다	매우 잘하는 편이다
1	국어	1	2	3	4	5
2	영어	1	2	3	4	5
3	수학	1	2	3	4	5
4	사회	1	2	3	4	5

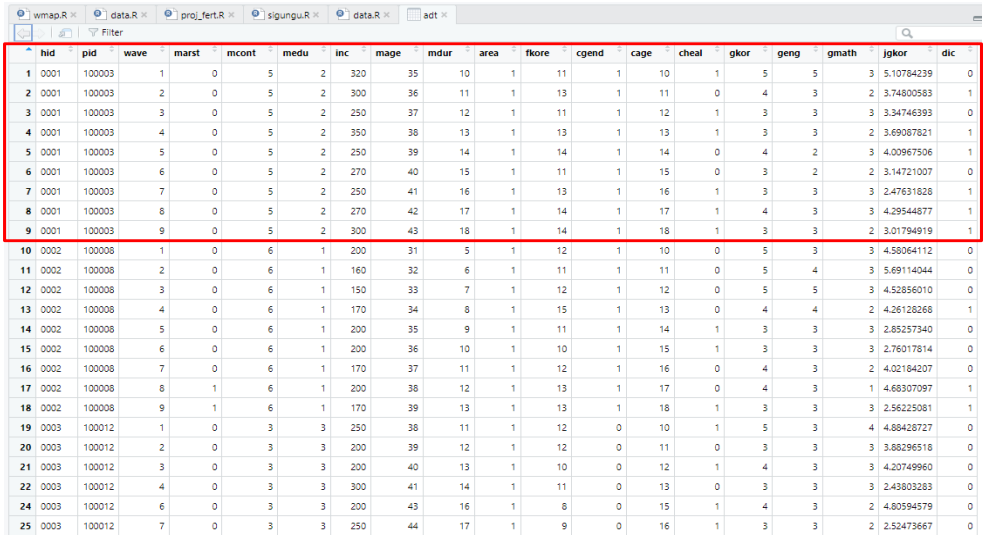
5

연구 가설 3

- 혼동 변수(confounding variables)
  - 원인 변수와 결과 변수 모두에 영향을 주는 변수로 이 변수들을 통제하지 않으면 원인 변수가 결과 변수에 주는 영향이 왜곡(biased)되어 나타남
  - 어머니 변수: 혼인상태, 출신국가, 교육수준, 월평균 소득, 연령, 한국에 산 기간, 거주지역
  - 아동 변수: 성별, 건강상태

6

종단 자료 만들기 1



wmap.R | data.R | proj_fert.R | sigungu.R | data.R | adt

Filter

	hid	pid	wave	marst	mcont	medu	inc	mage	mdur	area	fkore	cgend	cage	cheal	gkor	geng	gmath	jgkor	dic
1	0001	100003	1	0	5	2	320	35	10	1	11	1	10	1	5	5	3	5.10784239	0
2	0001	100003	2	0	5	2	300	36	11	1	13	1	11	0	4	3	2	3.74800583	1
3	0001	100003	3	0	5	2	250	37	12	1	11	1	12	1	3	3	3	3.34746393	0
4	0001	100003	4	0	5	2	350	38	13	1	13	1	13	1	3	3	2	3.69087821	1
5	0001	100003	5	0	5	2	250	39	14	1	14	1	14	0	4	2	3	4.00967506	1
6	0001	100003	6	0	5	2	270	40	15	1	11	1	15	0	3	2	2	3.14721007	0
7	0001	100003	7	0	5	2	250	41	16	1	13	1	16	1	3	3	3	2.47631828	1
8	0001	100003	8	0	5	2	270	42	17	1	14	1	17	1	4	3	3	4.29544877	1
9	0001	100003	9	0	5	2	300	43	18	1	14	1	18	1	3	3	2	3.01794919	1
10	0002	100008	1	0	6	1	200	31	5	1	12	1	10	0	5	3	3	4.58064112	0
11	0002	100008	2	0	6	1	160	32	6	1	11	1	11	0	5	4	3	5.69114044	0
12	0002	100008	3	0	6	1	150	33	7	1	12	1	12	0	5	5	3	4.52856010	0
13	0002	100008	4	0	6	1	170	34	8	1	15	1	13	0	4	4	2	4.26128268	1
14	0002	100008	5	0	6	1	200	35	9	1	11	1	14	1	3	3	3	2.85257340	0
15	0002	100008	6	0	6	1	200	36	10	1	10	1	15	1	3	3	3	2.76017814	0
16	0002	100008	7	0	6	1	170	37	11	1	12	1	16	0	4	3	2	4.02184207	0
17	0002	100008	8	1	6	1	200	38	12	1	13	1	17	0	4	3	1	4.68307097	1
18	0002	100008	9	1	6	1	170	39	13	1	13	1	18	1	3	3	3	2.56225081	1
19	0003	100012	1	0	3	3	250	38	11	1	12	0	10	1	5	3	4	4.88428727	0
20	0003	100012	2	0	3	3	200	39	12	1	12	0	11	0	3	3	3	3.88296516	0
21	0003	100012	3	0	3	3	200	40	13	1	10	0	12	1	4	3	3	4.20749960	0
22	0003	100012	4	0	3	3	300	41	14	1	11	0	13	0	3	3	3	2.43803283	0
24	0003	100012	6	0	3	3	200	43	16	1	8	0	15	1	4	3	2	4.80594579	0
25	0003	100012	7	0	3	3	250	44	17	1	9	0	16	1	3	3	2	2.52473667	0

7

종단 자료 만들기 2

- 순서
  - Loop start:
  - 각 년도의 자료에 대해
    - 원자료 불러들이기
    - 변수 이름 바꾸어 통일하기
    - 필요한 변수 추출
    - 각 년도 자료를 [rbind] 명령어를 사용하여 행으로 더하기
  - Loop end
  - 변수 재부호화 및 정리
  - 분석 변수 추출
  - listwise deletion
  - 분석 자료 저장

8

종단 자료 만들기 3

```
library(foreign)

# mother dataset
setwd("C:\\hyun\\hyun\\data\\maps\\다문화청소년패널조사(1기)_spss\\1기\\학부모\\")

for (i in 1:9) {
 eval(parse(text=paste("dt <- read.spss('다문화청소년패널 1기패널 학부모 ", i,"차년도.sav',
                        to.data.frame=TRUE, use.value.labels=FALSE)",  sep="")))
 eval(parse(text=paste("spdt <- subset(dt, select=c(ID, WAVE_w",i,
                ", family_a01_w", i, ", par_nat_1_w", i,
                ", par_edu_1_w", i, ", income_01_w", i,
                ", par_age_1_w", i,
                ", ko_stay_yr01_w", i,
                ", S_AREA1_w",i,
                ", ko_language_b1_w", i, ", ko_language_b2_w", i,
                ", ko_language_b3_w", i, ", ko_language_b4_w",i,
                "), subset=SURVEY3_w",i,"==1)",
                sep="")))
 names(spdt) <- c("hid", "wave", "family_a01", "par_nat_1",
                  "par_edu_1",  "income_01",
                  "par_age_1",
                  "ko_stay_yr01"
        ,         "s_area1",
                  "ko_language_b1", "ko_language_b2",
                  "ko_language_b3", "ko_language_b4")
 if (i==1) pdt <- spdt
 else pdt <- rbind(pdt, spdt)
}
pdt$hid <- gsub(" ","", pdt$hid)
pdt <- pdt[order(pdt$hid, pdt$wave),]
```

9

종단 자료 만들기 3

```
# child dataset
setwd("C:\\hyun\\hyun\\data\\maps\\다문화청소년패널조사(1기)_spss\\1기\\청소년\\")

for (i in 1:9) {
 eval(parse(text=paste("dt <- read.spss('다문화청소년패널 1기패널 청소년 ",
                  i,"차년도.sav', to.data.frame=TRUE, use.value.labels=FALSE)",
                  sep="")))
 eval(parse(text=paste("scdt <- subset(dt, select=c(ID, PID, WAVE_w",i,
                  ", S_GENDER_w", i, ", S_AGE_w", i, ", health_01_w",i,
                  ", score_01_w", i, ", score_02_w", i, ", score_03_w", i,
                  "), subset=SURVEY1_w",i,"==1)",
                  sep="")))
 names(scdt) <- c("hid", "pid", "wave",
                  "s_gender", "s_age", "health_01",
                  "score_01", "score_02", "score_03")
 if (i==1) cdt <- scdt
 else cdt <- rbind(cdt, scdt)
}
cdt <- cdt[order(cdt$hid, cdt$pid, cdt$wave),]
cdt$hid <- gsub(" ", "", cdt$hid)
```

10

종단 자료 만들기 4

```
# merge two datasets
adt <- merge(pdt, cdt, by=c("hid","wave"))
adt <- adt[order(adt$hid, adt$pid, adt$wave),]

# recoding
#marital status
adt$marst <- c(0,1,1,2,0)[match(adt$family_a01, 1:5)]
table(adt$family_a01, adt$marst, useNA="ifany")

# country of origin
adt$mcont <- adt$par_nat_1-2
table(adt$par_nat_1, adt$mcont, useNA="ifany")

# mother's education
adt$medu <- c(0,1,2,3,3)[match(adt$par_edu_1, 1:5)]
table(adt$par_edu_1, adt$medu, useNA="ifany")

# income
adt$inc <- adt$income_01

# mother's age
adt$mage <- adt$par_age_1

# duration in Korea in years
adt$mdur <- adt$ko_stay_yr01

# resident area
adt$area <- adt$s_area1-1

# fluency of Korean
adt$fkore <- adt$ko_language_b1 +
  adt$ko_language_b2 +
  adt$ko_language_b3 + adt$ko_language_b4

# child's variables
# gender
adt$cgend <- adt$s_gender-1

# agea
dt$cage <- adt$s_age

# health
adt$cheal <- adt$health_01-1

# self-reported grade
adt$gkor <- adt$score_01
adt$geng <- adt$score_02
adt$gmath <- adt$score_03

adt <- subset(adt,
        select=c(hid, pid, wave,
              marst, mcont, medu, inc, mage, mdur,
area, fkore,
              cgend, cage, cheal, gkor, geng, gmath))

# listwise deletion
adt <- adt[complete.cases(adt),]

write.csv(adt,
"C:\\hyun\\hyun\\2023_spring\\다문화청소년패널조사\\data\\adt.csv")
```

11

2: 기초 분석

- 기초 분석

12

기초 분석 1

- 평균과 다섯 요약값 보기
  - summary(adt)
    - 논리적으로 가능하지 않은 값은 없는가?
    - 결측값은 없는가?
    - 변수들의 분포는 적절한가?

```
> summary(adt)
     hid                pid              wave            marst        
 Length:11547       Min.   :100003   Min.   :1.000   Min.   :0.00000  
 Class :character   1st Qu.:101924   1st Qu.:2.000   1st Qu.:0.00000  
 Mode  :character   Median :103849   Median :5.000   Median :0.00000  
                    Mean   :103805   Mean   :4.705   Mean   :0.09639  
                    3rd Qu.:105649   3rd Qu.:7.000   3rd Qu.:0.00000  
                    Max.   :107560   Max.   :9.000   Max.   :2.00000  
     mcont             medu             inc              mage            mdur       
 Min.   :0.000   Min.   :0.000   Min.   :   0.0   Min.   :20.00   Min.   : 0.00  
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 150.0   1st Qu.:40.00   1st Qu.:13.00  
 Median :3.000   Median :1.000   Median : 220.0   Median :44.00   Median :16.00  
 Mean   :2.968   Mean   :1.472   Mean   : 244.9   Mean   :44.22   Mean   :15.81  
 3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.: 300.0   3rd Qu.:48.00   3rd Qu.:19.00  
 Max.   :6.000   Max.   :3.000   Max.   :2800.0   Max.   :68.00   Max.   :32.00  
      area            fkore            cgend             cage           cheal       
 Min.   :0.000   Min.   : 4.00   Min.   :0.0000   Min.   : 9.00   Min.   :0.0000  
 1st Qu.:1.000   1st Qu.:11.00   1st Qu.:0.0000   1st Qu.:11.00   1st Qu.:0.0000  
 Median :2.000   Median :12.00   Median :1.0000   Median :14.00   Median :1.0000  
 Mean   :2.207   Mean   :12.28   Mean   :0.5081   Mean   :13.67   Mean   :0.6723  
 3rd Qu.:3.000   3rd Qu.:13.00   3rd Qu.:1.0000   3rd Qu.:16.00   3rd Qu.:1.0000  
 Max.   :4.000   Max.   :16.00   Max.   :1.0000   Max.   :20.00   Max.   :3.0000  
      gkor            geng            gmath            jgkor             dic        
 Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :-0.1592   Min.   :0.0000  
 1st Qu.:3.000   1st Qu.:2.00   1st Qu.:2.000   1st Qu.: 2.8856   1st Qu.:0.0000  
 Median :3.000   Median :3.00   Median :3.000   Median : 3.5083   Median :0.0000  
 Mean   :3.513   Mean   :2.96   Mean   :3.002   Mean   : 3.5168   Mean   :0.3017  
 3rd Qu.:4.000   3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.: 4.1685   3rd Qu.:1.0000  
 Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   : 6.3515   Max.   :1.0000  
```

13

기초 분석 2

```
# categorize the causal variable
adt$dic <- ifelse(adt$fkore<=12,0,1)
table(adt$fkore, adt$dic, useNA="ifany")

vcat <- c(1,1,1,1,0,0,
          0,1,0,1,0,
          0,0,0,0)

for (i in 3:17) {
  vn <- names(adt)[i]
  if (vcat[i-2]==1) {
    eval(parse(text=paste("tb <- table(adt$",vn,")", sep="")))
    eval(parse(text=paste("tb <- cbind(tb, prop.table(table(adt$",
                          vn, "))*(-100))", sep="")))
    eval(parse(text=paste("tb <- cbind(tb, table(adt$",
                          vn,"[adt$dic==0]))", sep="")))
    eval(parse(text=paste("tb <- cbind(tb, prop.table(table(adt$",
                          vn, "[adt$dic==0]))*(-100))", sep="")))
    eval(parse(text=paste("tb <- cbind(tb, table(adt$",
                          vn, "[adt$dic==1]))", sep="")))
    eval(parse(text=paste("tb <- cbind(tb, prop.table(table(adt$",
                          vn, "[adt$dic==1]))*(-100))", sep="")))
  }
  else  eval(parse(text=paste("tb <- c(mean(adt$",
                              vn,"),-sd(adt$",vn,"), mean(adt$",
                              vn,"[adt$dic==0]),-sd(adt$",
                              vn, "[adt$dic==0]), mean(adt$",
                              vn,"[adt$dic==1]),-sd(adt$",vn,"[adt$dic==1]))",
                              sep="")))
  if (i==3) dsc <- tb
  else dsc <- rbind(dsc, tb)
}
write.csv(dsc, "C:\\hyun\\hyun\\2023_spring\\다문화청소년패널조사\\data\\dsc.csv")
```

14

기초 분석 3

		전체		낮은 한국어 능력		높은 한국어 능력	
		N/M	P/SD	N/M	P/SD	N/M	P/SD
전체		11,547	(100.0)	8,063	(69.8)	3,484	(30.2)
년도별 응답자 수	1	1,560	(13.5)	1,086	(13.5)	474	(13.6)
	2	1,437	(12.4)	1,002	(12.4)	435	(12.5)
	3	1,373	(11.9)	957	(11.9)	416	(11.9)
	4	1,309	(11.3)	900	(11.2)	409	(11.7)
	5	1,266	(11.0)	894	(11.1)	372	(10.7)
	6	1,247	(10.8)	888	(11.0)	359	(10.3)
	7	1,175	(10.2)	850	(10.5)	325	(9.3)
	8	1,116	(9.7)	741	(9.2)	375	(10.8)
	9	1,064	(9.2)	745	(9.2)	319	(9.2)
어머니 변수							
혼인상태	결혼동거	10,783	(93.4)	7,523	(93.3)	3,260	(93.6)
	이혼별거	415	(3.6)	301	(3.7)	114	(3.3)
	사별	349	(3.0)	239	(3.0)	110	(3.2)
출신국가	중국(조선족외)	847	(7.3)	510	(6.3)	337	(9.7)
	중국(조선족)	2,212	(19.2)	827	(10.3)	1,385	(39.8)
	베트남	276	(2.4)	230	(2.9)	46	(1.3)
	필리핀	3,020	(26.2)	2,634	(32.7)	386	(11.1)
	일본	4,128	(35.7)	2,985	(37.0)	1,143	(32.8)
	태국	446	(3.9)	385	(4.8)	61	(1.8)
	기타	618	(5.4)	492	(6.1)	126	(3.6)
교육수준	중졸이하	1,265	(11.0)	779	(9.7)	486	(13.9)
	고졸	5,431	(47.0)	3,755	(46.6)	1,676	(48.1)
	대졸(2년제)	2,990	(25.9)	2,196	(27.2)	794	(22.8)
	대졸(4년제 이상)	1,861	(16.1)	1,333	(16.5)	528	(15.2)
월평균소득		244.9	(119.9)	233.6	(109.0)	271.0	(138.6)
연령		44.2	(5.8)	44.4	(5.9)	43.9	(5.4)
한국에 산 기간		15.8	(4.2)	15.5	(4.2)	16.6	(4.1)
거주지역	서울	1,126	(9.8)	795	(9.9)	331	(9.5)
	경기 및 인천	2,991	(25.9)	2,114	(26.2)	877	(25.2)
	충청 및 강원	2,274	(19.7)	1,627	(20.2)	647	(18.6)
	경상권	2,682	(23.2)	1,878	(23.3)	804	(23.1)
	전라 및 제주	2,474	(21.4)	1,649	(20.5)	825	(23.7)
한국어 능력		12.2	(2.2)	11.1	(1.2)	15.0	(1.2)
아동 변수							
성별	남자	5,680	(49.2)	4,051	(50.2)	1,629	(46.8)
	여자	5,867	(50.8)	4,012	(49.8)	1,855	(53.2)
연령		13.7	(2.6)	13.7	(2.6)	13.6	(2.6)
건강상태		0.7	(0.6)	0.7	(0.6)	0.6	(0.6)
국어성적		3.5	(0.9)	3.5	(0.8)	3.6	(0.9)
영어성적		3.0	(1.1)	2.9	(1.1)	3.0	(1.1)
수학성적		3.0	(1.1)	3.0	(1.1)	3.1	(1.2)

15

기초 분석 4

- 연도별 노동조합 변수 변화 보기
  - with(adt, addmargins(table(wave, dic)))
  - with(adt, addmargins(prop.table(table(wave, dic),1)))
  - 연도가 증가할수록 사례수가 감소하고, 한국어 능력에서 13점 이상인 어머니의 비율이 감소하는 것 같다: why?

```
> with(adt, addmargins(table(wave, dic)))
     dic
wave    0    1  Sum
  1  1086  474 1560
  2  1002  435 1437
  3   957  416 1373
  4   900  409 1309
  5   894  372 1266
  6   888  359 1247
  7   850  325 1175
  8   741  375 1116
  9   745  319 1064
  Sum 8063 3484 11547
```

```
> with(adt, addmargins(prop.table(table(wave, dic),1)))
     dic
wave         0         1       Sum
  1   0.6961538 0.3038462 1.0000000
  2   0.6972860 0.3027140 1.0000000
  3   0.6970138 0.3029862 1.0000000
  4   0.6875477 0.3124523 1.0000000
  5   0.7061611 0.2938389 1.0000000
  6   0.7121091 0.2878909 1.0000000
  7   0.7234043 0.2765957 1.0000000
  8   0.6639785 0.3360215 1.0000000
  9   0.7001880 0.2998120 1.0000000
  Sum 6.2838424 2.7161576 9.0000000
```

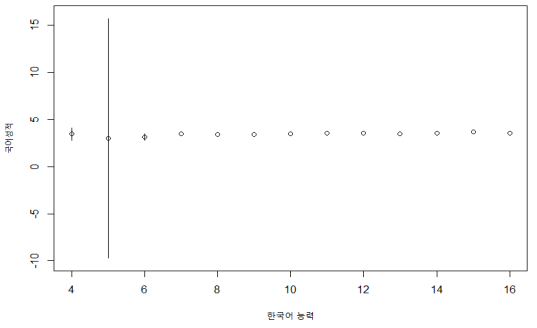
16

기초 분석 5

- 국어성적 평균과 95% 신뢰구간
  - t.test(adt$gkor)
  - t.test(adt$gkor)$estimate
  - t.test(adt$gkor)$conf.intf

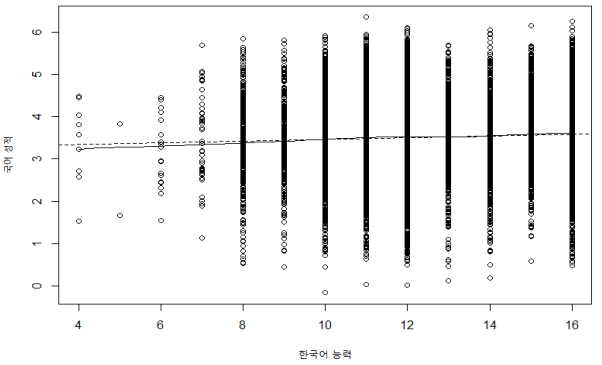
  - fciv <- function(x) c(t.test(x)$estimate,t.test(x)$conf.int[1:2])
  - civ <- tapply(adt$gkor, adt$fkore, fciv)
  - civ <- data.frame(cbind(4:16,do.call(rbind, civ)))
  - names(civ)[2] <- "V2"

  - plot(civ$V1, civ$V2, ylim=c(-10, 16), xlab="한국어 능력", ylab="국어성적")
  - for (i in 1:3) lines(c(civ$V1[i], civ$V1[i]), c(civ$V3[i], civ$V4[i]))



기초 분석 6

- 국어성적 산점도
- fit0 <- lm(gkor ~ fkore, data=adt)
- set.seed(6)
- adt$jgkor <- adt$gkor+rnorm(dim(adt)[1],mean=0,0.4)
- scatter.smooth(adt$fkore, adt$jgkor, xlab="한국어 능력",
- ylab="국어 성적")
- abline(fit0, lty=2)



8

3: 일반최소제곱법

- 일반최소제곱법
- Sandwich 추정법

19

일반최소제곱법 1

- 어떤 특정 변수의 값을 예측하고자 할 때 가장 많이 사용하는 추정법이 ordinary least squares (OLS:일반최소제곱법)이다
- 우리의 모형을 다음과 같이 나타낸다고 하자
- $y_i = \sum_k \beta_k X_{ki} + \epsilon_i$ 혹은 $Y = X^T\beta + \epsilon$
  where K는 모수의 개수(변수 수+1)이며 i는 개별 사례를 나타낸다
- OLS는 β를 다음과 같이 추정한다
- $\min_\beta \epsilon'\epsilon = \min_\beta [Y - X^T\beta]^T[Y - X^T\beta]$
  - $\hat{\beta} = (X^TX)^{-1}X^TY$
  - $\widehat{var}(\hat{\beta}) = var[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^T var(\epsilon)X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \frac{1}{N-K}(\hat{\epsilon}^T\hat{\epsilon})(X^TX)^{-1}$

20

일반최소제곱법 2

- Gauss-Markov Theorem에 따르면 다음의 조건 하에 위의 추정치가 BLUE(best linear unbiased estimator)라는 것을 보여준다
  - best는 가장 효율적이라는 의미이다
- 조건 1: $E(\epsilon_i) = 0$, 모든 i에 대해
- 조건 2: $var(\epsilon_i) = \sigma^2$, 모든 i에 대해
  - 동분산성(homoskedasticity) 가정
- 조건 3: $cov(\epsilon_i, \epsilon_j) = 0$, 모든 $i \neq j$에 대해
  - 잔차의 독립성(independence) 가정
- 조건 4: $cov(x_{ki}, \epsilon_i) = 0$, 모든 k 및 i에 대해
  - 외생성(exogeneity) 가정

21

일반최소제곱법 3

- OLS에서 어떤 계수 $\hat{\beta}_k$의 해석
  - 다른 변수를 통제한 상태에서 X_k가 한 단위 증가할 때 Y의 변화량
- 특정 모형의 예측력이 얼마나 좋은가를 평가하는 하나의 통계치로 때때로 다중결정계수(coefficient of multiple determination)라 불리는 R^2가 쓰인다
  - $R^2 = \frac{TSS-SSE}{TSS} = \frac{\sum(Y-\bar{Y})^2-\sum(Y-\hat{Y})^2}{\sum(Y-\bar{Y})^2} = corr(Y, \hat{Y})^2$
- t-test:
  - $H_0: \beta_k = 0, H_A: \beta_k \neq 0$
  - $t = \frac{\hat{\beta}_k}{se}$ 를 사용하며 이 통계치는 $df = N - K$인 t-분포를 따른다
- F-test
  - H_0: $\beta_2 = \cdots = \beta_K = 0, (\beta_1\ is\ for\ the\ intercept), H_A$: at least one β is not zero
  - $F = \frac{R^2/(K-1)}{(1-R^2)/[N-K]}$ 값은 $df_1 = K - 1$ 이고 $df_2 = N - K$인 F-분포를 따른다

22

일반최소제곱법 4

- $grade = \beta_0 + \beta_1 fkore + X^T\alpha + \epsilon$
  - fit1 <- lm(gkor ~ factor(marst) +factor(mcont) +
  - factor(medu) + inc + mage + mdur +
  - factor(area) + fkore +
  - factor(cgend) + cage + cheal,
  - data=adt)
  - summary(fit1)

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.968e+00  9.626e-02  41.225  < 2e-16 ***
factor(marst)1 -4.276e-03  4.255e-02  -0.100  0.91996
factor(marst)2 -7.067e-02  4.588e-02  -1.540  0.12348
factor(mcont)1 -5.951e-02  3.385e-02  -1.758  0.07876 .
factor(mcont)2  5.042e-02  5.873e-02   0.859  0.39062
factor(mcont)3 -4.582e-02  3.396e-02  -1.349  0.17735
factor(mcont)4 -1.635e-02  3.370e-02  -0.485  0.62767
factor(mcont)5 -6.519e-03  4.881e-02  -0.134  0.89376
factor(mcont)6 -6.270e-02  4.465e-02  -1.404  0.16025
factor(medu)1   4.334e-02  2.710e-02   1.599  0.10976
factor(medu)2   9.747e-02  3.079e-02   3.165  0.00155 **
factor(medu)3   1.909e-01  3.248e-02   5.877 4.29e-09 ***
inc             1.074e-04  7.121e-05   1.508  0.13160
mage            5.432e-03  1.873e-03   2.901  0.00373 **
mdur           -1.487e-02  2.733e-03  -5.442 5.38e-08 ***
factor(area)1  -2.647e-02  2.900e-02  -0.913  0.36145
factor(area)2   3.362e-02  3.038e-02   1.106  0.26855
factor(area)3   2.277e-02  2.970e-02   0.767  0.44323
factor(area)4  -1.740e-02  3.017e-02  -0.577  0.56412
fkore           2.757e-02  4.213e-03   6.545 6.18e-11 ***
factor(cgend)1  1.591e-01  1.542e-02  10.317  < 2e-16 ***
cage           -6.323e-02  3.993e-03 -15.835  < 2e-16 ***
cheal          -1.273e-01  1.226e-02 -10.378  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8218 on 11524 degrees of freedom
Multiple R-squared:  0.0781,    Adjusted R-squared:  0.07634
F-statistic: 44.37 on 22 and 11524 DF,  p-value: < 2.2e-16
```


Sandwich 추정법 I

- 앞서 $var(\epsilon) = \sigma^2$로 가정하였으며 이는 조건 2에 의한 것이다.
- $\widehat{var}(\hat{\beta}) = (X^TX)^{-1}X^T var(\epsilon)X(X^TX)^{-1}$
- 하지만 패널 자료는 한 사례당 여러 년도의 관측값이 있기 때문에 사례에 따라 다른 분산을 가지고 있을 가능성이 높다
  - 어떤 아이들은 성적의 변동이 클 것이고, 어떤 아이들은 성적의 변동이 적을 것이다
- 잔차 가정이 만족스럽지 못하다면 하나의 대안은 Sandwich 추정법으로 분산을 구하는 것이다
  - 흔히 $(X^TX)^{-1}$를 bread part라고 하고 $X^T var(\epsilon)X$를 meat part라고 한다
  - 또한 많은 경우 $var(\epsilon)$를 Ω로 표시한다

Sandwich 추정법 2

- 모든 잔차 사이에 이분산성(heteroskedasticity)이 있으면
  - $X^T \widehat{var}(\epsilon) X$에서 $\widehat{var}(\epsilon) = \hat{\Omega} = \frac{N}{N-K}\text{diag}(\hat{\epsilon}_1, \hat{\epsilon}_2, \ldots, \hat{\epsilon}_N)$ 으로 추정하고 $\hat{\epsilon}$는 추정된 잔차임
- 만약 집단 내 상관관계를 고려하고 싶다면
  - $X^T \widehat{var}(\epsilon) X = \sum_{g=1}^{G} \hat{u}_g^T \hat{u}_j, \hat{u}_j = \sum_{i=1}^{N_G} \hat{\epsilon}_i * x_i$
  - 표본의 수가 무한대일 수 없다는 점을 교정하기 위해 (finite-sample adjustment) <식 4.5>에 $(N-1)G/[(N-K)(G-1)]$를 곱해준다

25

Sandwich 추정법 3

- 통상의 추정법
  - fit1 <- lm(gkor ~ factor(marst) +factor(mcont) +
  - factor(medu) + inc + mage + mdur +
  - factor(area) + fkore +
  - factor(cgend) + cage + cheal,
  - data=adt)
- Sandwich 추정법
  - #install.packages(sandwich)
  - library(sandwich)
  - library(lmtest)
  - coeftest(fit1, df=Inf, vcov=vcovHC(fit1, type= " HC1 "))
  - Stata에서는 vce(robust) 옵션과 같음
- Clustered sandwich 추정법
  - #install.packages("multiwayvcov")
  - library(multiwayvcov)
  - fit1_clv <- cluster.vcov(fit1, adt$pid)
  - coeftest(fit1, vcov = fit1_clv)
  - Stata에서는 vce(cluster pid) 옵션과 같음

26

Sandwich 추정법 4

OLS

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.968e+00  9.626e-02  41.225  < 2e-16 ***
factor(marst)1 -4.276e-03  4.255e-02  -0.100  0.91996
factor(marst)2 -7.067e-02  4.588e-02  -1.540  0.12348
factor(mcont)1 -5.951e-02  3.385e-02  -1.758  0.07876 .
factor(mcont)2  5.042e-02  5.873e-02   0.859  0.39062
factor(mcont)3 -4.582e-02  3.396e-02  -1.349  0.17735
factor(mcont)4 -1.635e-02  3.370e-02  -0.485  0.62767
factor(mcont)5 -6.519e-03  4.881e-02  -0.134  0.89376
factor(mcont)6 -6.270e-02  4.465e-02  -1.404  0.16025
factor(medu)1   4.334e-02  2.710e-02   1.599  0.10976
factor(medu)2   9.747e-02  3.079e-02   3.165  0.00155 **
factor(medu)3   1.909e-01  3.248e-02   5.877 4.29e-09 ***
inc             1.074e-04  7.121e-05   1.508  0.13160
mage            5.432e-03  1.873e-03   2.901  0.00373 **
mdur           -1.487e-02  2.733e-03  -5.442 5.38e-08 ***
factor(area)1  -2.647e-02  2.900e-02  -0.913  0.36145
factor(area)2   3.362e-02  3.038e-02   1.106  0.26855
factor(area)3   2.277e-02  2.970e-02   0.767  0.44323
factor(area)4   1.749e-02  3.017e-02   0.577  0.56412
fkore           2.757e-02  4.213e-03   6.545 6.18e-11 ***
factor(cgend)1  1.591e-01  1.542e-02  10.317  < 2e-16 ***
cage           -6.323e-02  3.993e-03 -15.835  < 2e-16 ***
cheal          -1.273e-01  1.226e-02 -10.378  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8218 on 11524 degrees of freedom
Multiple R-squared:  0.0781,    Adjusted R-squared:  0.07634
F-statistic: 44.37 on 22 and 11524 DF,  p-value: < 2.2e-16
```

Sandwich

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.96829147	0.09738872	40.7469	< 2.2e-16	***
factor(marst)1	-0.00427579	0.03870208	-0.1105	0.912029	
factor(marst)2	-0.07067128	0.04431009	-1.5949	0.110729	
factor(mcont)1	-0.05951238	0.03407781	-1.7464	0.080747	.
factor(mcont)2	0.05042064	0.05868717	0.8591	0.390262	
factor(mcont)3	-0.04582006	0.03393771	-1.3501	0.176977	
factor(mcont)4	-0.01634704	0.03402085	-0.4805	0.630871	
factor(mcont)5	-0.00651910	0.04644182	-0.1404	0.888367	
factor(mcont)6	-0.06269683	0.04621376	-1.3567	0.174886	
factor(medu)1	0.04333871	0.02758779	1.5709	0.116197	
factor(medu)2	0.09746970	0.03090283	3.1541	0.001610	**
factor(medu)3	0.19088240	0.03304340	5.7767	7.617e-09	***
inc	0.00010737	0.00006889	1.5586	0.119085	
mage	0.00543151	0.00191155	2.8414	0.004491	**
mdur	-0.01487215	0.00284023	-5.2362	1.639e-07	***
factor(area)1	-0.02646605	0.03023450	-0.8754	0.381378	
factor(area)2	0.03361873	0.03137985	1.0713	0.284013	
factor(area)3	0.02277108	0.03087076	0.7376	0.460742	
factor(area)4	-0.01739867	0.03140827	-0.5541	0.579627	
fkore	0.02757336	0.00434483	6.3462	2.206e-10	***
factor(cgend)1	0.15911475	0.01540672	10.3276	< 2.2e-16	***
cage	-0.06322749	0.00406435	-15.5566	< 2.2e-16	***
cheal	-0.12727539	0.01296480	-9.8170	< 2.2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Clustered sandwich

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.96829146	0.15212602	26.0856	< 2.2e-16	***
factor(marst)1	-0.00427579	0.06065672	-0.0705	0.943804	
factor(marst)2	-0.07067128	0.07301463	-0.9679	0.333112	
factor(mcont)1	-0.05951238	0.06053745	-0.9831	0.325595	
factor(mcont)2	0.05042064	0.10694687	0.4715	0.637325	
factor(mcont)3	-0.04582006	0.05967940	-0.7678	0.442639	
factor(mcont)4	-0.01634704	0.06110331	-0.2675	0.789065	
factor(mcont)5	-0.00651911	0.07864106	-0.0829	0.933935	
factor(mcont)6	-0.06269683	0.07959770	-0.7877	0.430905	
factor(medu)1	0.04333871	0.04818409	0.8994	0.368437	
factor(medu)2	0.09746970	0.05403774	1.8037	0.071299	.
factor(medu)3	0.19088240	0.05934216	3.2166	0.001301	**
inc	0.00010737	0.00011039	0.9726	0.330753	
mage	0.00543151	0.00344758	1.5755	0.115179	
mdur	-0.01487215	0.00493919	-3.0110	0.002609	**
factor(area)1	-0.02646605	0.05506570	-0.4806	0.630791	
factor(area)2	0.03361873	0.05741898	0.5855	0.558224	
factor(area)3	0.02277108	0.05669559	0.4016	0.687958	
factor(area)4	-0.01739867	0.05870459	-0.2964	0.766948	
fkore	0.02757336	0.00603087	4.5720	4.880e-06	***
factor(cgend)1	0.15911475	0.02776640	5.7305	1.027e-08	***
cage	-0.06322749	0.00611322	-10.3427	< 2.2e-16	***
cheal	-0.12727539	0.01692652	-7.5193	5.916e-14	***

27

4: 고정효과 모형

- 기본 모형 및 추정방법
- within 추정
- LSDV 추정
- 1차 차분 추정

28

기본 모형

- 다음의 패널 선형회귀모형을 가정하자
- $y_{it} = \alpha + \beta x_{it} + u_i + e_{it}, \quad i = 1, \ldots, n$ 및 $t = 1, \ldots, T$
- 에서 오차항은 시간에 따라 변하지 않는 패널의 개체 특성을 나타내는 u_i 와 시간과 패널 개체에 따라 변하는 순수한 오차항인 e_{it}로 구성되어 있다
- 이제 오차항 u_i를 확률변수(random variable)가 아닌 추정해야 할 모수(parameter)로 간주하는 고정효과(fixed effects) 모형에 대해 알아보자
  - 다른 말로 하면 u_i가 어떤 특정한 분포를 따르지 않는다는 것이다
- 앞이 식을 달리 쓰면 $y_{it} = (\alpha + u_i) + \beta x_{it} + e_{it}$
  - 고정효과 모형은 상수항이 패널 개체 별로 서로 다르면서 고정되어(fixed) 있다고 가정한다

29

Within 추정 1

- within 추정법을 알아보기 위해 각 패널의 연도별 변수값에 대해 평균을 구한 between 모형을 적어보자
- $\bar{y}_i = \alpha + \beta \bar{x}_i + u_i + \bar{e}_i$
- 이제 앞선 식에서 between 모형을 빼 주면 (within 변환)
- $(y_{it} - \bar{y}_i) = (\alpha - \alpha) + \beta(x_{it} - \bar{x}_i) + (u_i - u_i) + (e_{it} - \bar{e}_i) = \beta(x_{it} - \bar{x}_i) + (e_{it} - \bar{e}_i)$
- α와 u_i가 사라진 것을 알 수 있다 (differenced out). 따라서 $cov(x_{it}, u_i) \neq 0$ 라고 하더라도 OLS 추정을 통해 β에 대한 일치추정량을 구할 수 있다
- 위 식를 추정한 후 u_i를 사후적으로 계산할 수 있다

30

R

Within 추정 2

- install.packages("plm")
- library(plm)
- padt <- pdata.frame(adt, index=c("pid","wave"), drop.index=TRUE)
- fit2 <- plm(gkor ~ factor(marst) +factor(mcont) + factor(medu) +
- inc + mage + mdur + factor(area) + fkore +
- factor(cgend) + cage + cheal,
- data=padt, model="within")
- summary(fit2)

```
Coefficients: (1 dropped because of singularities)
                 Estimate  Std. Error t-value  Pr(>|t|)
factor(marst)1 -0.04892119  0.09283463 -0.5270   0.59823
factor(marst)2  0.02575830  0.09118361  0.2825   0.77758
factor(medu)1   0.14470868  0.26011505  0.5563   0.57800
factor(medu)2   0.56027968  0.71381138  0.7849   0.43252
factor(medu)3   0.84597471  0.54833673  1.5428   0.12291
inc            -0.00012309  0.00010301 -1.1949   0.23216
mage           -0.04513623  0.02700311 -1.6715   0.09465 .
mdur           -0.02611671  0.02720662 -0.9599   0.33711
factor(area)1  -0.06476419  0.14924316 -0.4340   0.66433
factor(area)2  -0.20744239  0.23461700 -0.8842   0.37662
factor(area)3   0.13262154  0.27810380  0.4769   0.63346
factor(area)4   0.07865720  0.22602699  0.3480   0.72785
fkore           0.01050642  0.00494063  2.1265   0.03348 *
cheal          -0.08534958  0.01275227 -6.6929 2.306e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

31

Stata

Within 추정 3

- set more off
- insheet using "C:\hyun\hyun\2023_spring\다문화청소년패널조사\data\adt.csv", clear
- xtset pid wave
- xi: xtreg gkor i.marst i.mcont i.medu inc mage mdur i.area ///
- fkore i.cgend cage cheal, fe

```
Fixed-effects (within) regression               Number of obs      =     11547
Group variable: pid                             Number of groups   =      1566

R-sq:  within  = 0.0797                         Obs per group: min =         1
       between = 0.0130                                        avg =       7.4
       overall = 0.0219                                        max =         9

                                                F(14,9967)         =     61.69
corr(u_i, Xb)  = -0.5028                        Prob > F           =    0.0000
```

gkor	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Imarst_1	-.0489212	.0928346	-0.53	0.598	-.2308958	.1330534
_Imarst_2	.0257583	.0911836	0.28	0.778	-.15298	.2044966
_Imcont_1	0	(omitted)				
_Imcont_2	0	(omitted)				
_Imcont_3	0	(omitted)				
_Imcont_4	0	(omitted)				
_Imcont_5	0	(omitted)				
_Imcont_6	0	(omitted)				
_Imedu_1	.1447087	.2601151	0.56	0.578	-.3651694	.6545867
_Imedu_2	.5602797	.7138114	0.78	0.433	-.8389348	1.959494
_Imedu_3	.8459747	.5483367	1.54	0.123	-.2288761	1.920825
inc	-.0001231	.000103	-1.19	0.232	-.000325	.0000788
mage	-.0451362	.0270031	-1.67	0.095	-.0980678	.0077953
mdur	-.0261167	.0272066	-0.96	0.337	-.0794472	.0272138
_Iarea_1	-.0647642	.1492432	-0.43	0.664	-.3573109	.2277826
_Iarea_2	-.2074424	.234617	-0.88	0.377	-.6673391	.2524543
_Iarea_3	.1326215	.2781038	0.48	0.633	-.4125181	.6777612
_Iarea_4	.0786572	.226027	0.35	0.728	-.3644014	.5217158
fkore	.0105064	.0049406	2.13	0.033	.0008218	.020191
_Icgend_1	0	(omitted)				
cage	0	(omitted)				
cheal	-.0853496	.0127523	-6.69	0.000	-.1103466	-.0603526
_cons	5.541821	.8469972	6.54	0.000	3.881536	7.202107
sigma_u	.6821432					
sigma_e	.68190044					
rho	.50017797	(fraction of variance due to u_i)				

F test that all u_i=0: F(1565, 9967) = 4.44 Prob > F = 0.0000

Stata

Within 추정 4

- Within 추정법을 사용하여 얻은 결과를 보면 [corr(u_i, Xb)]가 있다
- 이는 $corr\left(\hat{u}_i = \bar{y}_i - \hat{\alpha}_{FE} - \hat{\beta}_{FE}\bar{x}_i, \hat{y}_{it} = \hat{\alpha}_{FE} + \hat{\beta}_{FE}x_{it}\right)$이다
- 이 통계값은 $cov(x_{it}, u_i)$의 정도를 나타낸다
  - 이 값이 작으면 상관관계가 적은 것이므로 확률효과 모형을 선호한다
  - 이 값이 크면 확률효과 모형의 가정을 위배하는 것이므로 고정효과 모형을 선호한다

Stata

Within 추정 5

- [xtreg, fe]의 추정계수 다음에는 [sigma_u]와 [sigma_e], 그리고 [rho]가 제시되어 있다
- 이들은 각각 $\hat{\sigma}_u$, $\hat{\sigma}_e$, $\hat{\rho} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} = \frac{1}{1 + \hat{\sigma}_e^2/\hat{\sigma}_u^2}$을 나타낸다
- 달리 말하면 $\hat{\rho}$는 오차항의 총분산에서 패널의 개체특성을 의미하는 오차항(u_i)의 분산이 차지하는 비율을 뜻한다
  - $\hat{\rho}$는 항상 0과 1 사이의 값을 갖는다
  - $\hat{\rho}$의 값이 크면 클수록 독립변수에 의해 설명되지 않는 종속변수 변동의 많은 부분이 패널의 개체특성에 의해 설명된다는 의미이다
  - 즉 이 값이 1에 가까울수록 시간에 따라 변하지 않는 패널 개체의 특성을 감안하는 것이 중요하다는 것이다
- 결과의 마지막 줄에는 H_0: 모든 i에 대해 $u_i = 0$에 대한 검정결과를 제시하고 있다
  - 이 검정결과가 유의하지 않다면, 즉 유의도가 0.05보다 크다면 패널 개체의 이질성을 고려할 필요가 없다
  - F-검정결과 영가설이 기각되면 고정효과 모형을 선택한다

Stata

Within 추정 6

- $y_{it} = \alpha + \beta x_{it} + u_i + e_{it}$에서 $u_i = \tau z_i$로 볼 수 있다
- 이때 z_i는 시간불변인 패널 개체 별 특성, 예를 들어 IQ나 외모와 같은 것이며 τ는 이와 연관된 계수이다. 이런 면에서 u_i는 관찰되지 않은 시간불변 변수의 *시간불변* 효과를 통제하는 측면이 있다
- 헌데, u_i는 모형 추정과정에서 빠지므로 시간불변 변수의 효과를 추정하지 못한다는 약점이 있다
- 예를 들어, 성별이나 어머니의 국가와 같은 패널 내 불변변수의 효과는 추정할 수 없다

35

LSDV 추정 1

- 앞선 모형에 따르면 $y_{it} = (\alpha + u_i) + \beta x_{it} + e_{it}$
- u_i를 고정효과로 본다면 $(\alpha + u_i)$는 각 개체가 평균으로부터 얼마나 떨어져 있는가를 나타내는 것으로 볼 수 있다
- 이를 달리 모수화한다면 평균으로부터 떨어져 있는 것이 아닌 각 개체의 상수항을 추정하는 다음과 같은 모형으로 쓸 수 있다
- $y_{it} = \sum_{i=1}^{n} \alpha_i + \beta x_{it} + e_{it}$
- 즉 각 패널의 더미변수를 만들어 이를 OLS로 추정하면 고정효과 모형을 추정할 수 있다는 것이다
- 이러한 추정방법을 LSDV(least squares dummy variable) 추정법이라 한다

36

Stata

LSDV 추정 2

- 패널 수가 많으면 [reg]명령어를 사용했을 때 매우 많은 더미변수가 생겨 추정에 시간이 걸리며 결과도 지저분해 보인다
- 따라서 통상 [areg, ab()]라는 명령어를 사용한다
  - *xi: areg gkor i.marst i.mcont i.medu inc mage mdur i.area ///*
  - *fkore i.cgend cage cheal, ab(pid)*

```
Linear regression, absorbing indicators          Number of obs   =     11547
                                                 F(  14,  9967) =     61.69
                                                 Prob > F        =    0.0000
                                                 R-squared       =    0.4511
                                                 Adj R-squared   =    0.3641
                                                 Root MSE        =    0.6819
```

gkor	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Imarst_1	-.0489212	.0928346	-0.53	0.598	-.2308958	.1330534
_Imarst_2	.0257583	.0911836	0.28	0.778	-.15298	.2044966
_Imcont_1	0	(omitted)				
_Imcont_2	0	(omitted)				
_Imcont_3	0	(omitted)				
_Imcont_4	0	(omitted)				
_Imcont_5	0	(omitted)				
_Imcont_6	0	(omitted)				
_Imedu_1	.1447087	.2601151	0.56	0.578	-.3651694	.6545867
_Imedu_2	.5602797	.7138114	0.78	0.433	-.8389348	1.959494
_Imedu_3	.8459747	.5483367	1.54	0.123	-.2288761	1.920825
inc	-.0001231	.000103	-1.19	0.232	-.000325	.0000788
mage	-.0451362	.0270031	-1.67	0.095	-.0980678	.0077953
mdur	-.0261167	.0272066	-0.96	0.337	-.0794472	.0272138
_Iarea_1	-.0647642	.1492432	-0.43	0.664	-.3573109	.2277826
_Iarea_2	-.2074424	.234617	-0.88	0.377	-.6673391	.2524543
_Iarea_3	.1326215	.2781038	0.48	0.633	-.4125181	.6777612
_Iarea_4	.0786572	.226027	0.35	0.728	-.3644014	.5217158
fkore	.0105064	.0049406	2.13	0.033	.0008218	.020191
_Icgend_1	0	(omitted)				
cage	0	(omitted)				
cheal	-.0853496	.0127523	-6.69	0.000	-.1103466	-.0603526
_cons	5.541821	.8469972	6.54	0.000	3.881536	7.202107
pid	F(1565, 9967) =	4.690	0.000		(1566 categories)	

37

1차 차분 추정 1

- 이제 세 번의 시기에 걸친 자료가 있다고 하자
- $y_{i1} = \alpha + \beta x_{i1} + u_i + e_{i1}, \quad t = 1$
- $y_{i2} = \alpha + \beta x_{i2} + u_i + e_{i2}, \quad t = 2$
- $y_{i3} = \alpha + \beta x_{i3} + u_i + e_{i3}, \quad t = 3$
- 1차 차분을 실행하면
- $\Delta y_{i2} = \Delta\beta x_{i2} + \Delta e_{i2}, \quad t = 2$
- $\Delta y_{i3} = \Delta\beta x_{i3} + \Delta e_{i3}, \quad t = 3$
- $T = 2$인 경우에는 1차 차분을 실행하면 패널자료가 횡단면자료로 전환되지만 $T > 2$이면 1차 차분을 해도 패널자료 구조를 유지하게 된다
- 1차 차분을 합동한 다음 OLS를 적용하면
  - u_i가 없어져서 $corr(x_{it}, u_i) \neq 0$이더라도 일치추정량을 구할 수 있다
  - 그러나 오차항 e_{it}에 자기상관이 존재하지 않더라도 1차 차분 모형의 오차항 Δe_{it}에는 1계 자기상관이 존재한다. 즉 $cov(\Delta e_{it}, \Delta e_{it-1}) = cov(e_{it} - e_{it-1}, e_{it-1} - e_{it-2}) = -\sigma_e^2$
  - 오히려 e_{it}에 $e_{it} = e_{it-1} + v_{it}$와 같이 1계 자기상관이 있으면, $cov(\Delta e_{it}, \Delta e_{it-1}) = cov(v_{it}, v_{it-1}) = 0$
  - 다시 말해 e_{it}에 계수가 1인 1계 자기상관이 있어야만 효율적 추정량을 얻을 수 있다

38

R

1차 차분 추정 2

```
fit3 <- plm(gkor ~ -1 + factor(marst) +factor(mcont) +
            factor(medu) + inc + mage + mdur +
            factor(area) + fkore + factor(cgend) +
           cage + cheal,   data=padt, model="fd")
summary(fit3)
```

```
Unbalanced Panel: n = 1566, T = 1-9, N = 11547
Observations used in estimation: 9981

Residuals:
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-4.0571 -0.8722  0.0569 -0.0018  0.1428  4.0602

Coefficients: (2 dropped because of singularities)
                 Estimate  Std. Error t-value  Pr(>|t|)
factor(marst)0  0.00043626  0.14337133  0.0030   0.99757
factor(marst)1  0.07969856  0.18953418  0.4205   0.67413
factor(medu)1  -0.45806060  0.44033785 -1.0402   0.29825
factor(medu)2  -1.39817482  1.32079021 -1.0586   0.28981
factor(medu)3  -0.33963272  0.98461031 -0.3449   0.73015
inc            -0.00018719  0.00011984 -1.5619   0.11833
mage            0.01872461  0.03286494  0.5697   0.56886
mdur           -0.07558733  0.03257377 -2.3205   0.02033 *
factor(area)1   0.16671533  0.23143323  0.7204   0.47132
factor(area)2  -0.07292161  0.33966152 -0.2147   0.83001
factor(area)3   0.32834349  0.38375752  0.8556   0.39224
factor(area)4   0.79331751  0.34253748  2.3160   0.02058 *
fkore           0.00167938  0.00479218  0.3504   0.72601
cheal          -0.06718081  0.01250304 -5.3732 7.911e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    7763.5
Residual Sum of Squares: 7726.8
R-Squared:      0.0047358
Adj. R-Squared: 0.0034376
F-statistic: 6.65004 on 14 and 9967 DF, p-value: 1.1575e-13
```


Stata

1차 차분 추정 3

- reg D.gkor D.(inc mage mdur fkore cage cheal _I*), nocon
- 두 프로그램의 계수와 표준오차가 약간 다르다
- e_{it}에 자기상관이 없으면 within 추정은 적절하면 1차 차분 모형은 효율적이지 않다

- 하지만 둘 다 일치추정량이다
- 반대로 e_{it}가 1계 자기상관이 있으면 1차 차분 모형이 적절하며 within 추정이 효율적이지 않다
- Δe_{it}에 1계 자기상관이 존재하면 clustered Sandwich 추정량을 이용하여 오차항끼리 독립이라는 가정을 완화한다
- reg D.gkor D.(inc mage mdur fkore cage cheal _I*), nocon vce(cluster pid)

Source	SS	df	MS	Number of obs = 9911
Model	73.9625217	14	5.2830944	F(14, 9997) = 6.83
Residual	7646.13748	9997	.772871232	Prob > F = 0.0000
				R-squared = 0.0096
Total	7720	9911	.778932499	Adj R-squared = 0.0082
				Root MSE = .87934

D.gkor	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
inc						
D1.	-.0001657	.0001206	-1.37	0.169	-.000402	.0000707
mage						
D1.	.0106725	.033781	0.32	0.752	-.0555452	.0768901
mdur						
D1.	-.0706728	.0334848	-2.11	0.035	-.1363095	-.0050361
fkore						
D1.	.0017722	.004797	0.37	0.712	-.0076309	.0111753
cage						
D1.	0	(omitted)				
cheal						
D1.	-.0682074	.0125255	-5.45	0.000	-.0927599	-.0436549
_Imarst_1						
D1.	.1007324	.1345243	0.75	0.454	-.1629627	.3644274
_Imarst_3						
D1.	-.0300798	.145015	-0.21	0.836	-.3143387	.2541792
_Imcont_1						
D1.	0	(omitted)				
_Imcont_2						
D1.	0	(omitted)				
_Imcont_3						
D1.	0	(omitted)				
_Imcont_4						
D1.	0	(omitted)				
_Imcont_5						
D1.	0	(omitted)				
_Imcont_6						
D1.	0	(omitted)				
_Imedu_1						
D1.	-.4583948	.439583	-1.04	0.300	-1.317047	.4062774
_Imedu_2						
D1.	-1.39393	1.318516	-1.06	0.290	-3.97849	1.190631
_Imedu_3						
D1.	-.3321571	.982922	-0.34	0.735	-2.258854	1.59457
_Iarea_1						
D1.	.1474221	.2310365	0.72	0.469	-.2854566	.6203008
_Iarea_2						
D1.	-.0721796	.3390794	-0.21	0.831	-.7368442	.5924851
_Iarea_3						
D1.	.3296034	.3831031	0.86	0.390	-.4212746	1.080443
_Iarea_4						
D1.	.7946065	.3419522	2.32	0.020	.1243906	1.464823
_Icgend_1						
D1.	0	(omitted)				

Stata

1차 차분 추정 4

- 1차 차분 모형에서 가장 중요한 가정은 e_{it}에 자기상관이 존재하는가의 여부이므로 이를 검정할 필요가 있다
- [xtserial] 명령어는 Wooldridge가 제안한 방법을 명령어로 만들어 놓았다
- e_{it}에 자기상관이 없다는 영가설 하에 $corr(\Delta e_{it}, \Delta e_{it-1}) = \frac{cov(\Delta e_{it}, \Delta e_{it-1})}{\sqrt{var(\Delta e_{it})}\sqrt{var(\Delta e_{it-1})}} = \frac{cov(e_{it}-e_{it-1}, e_{it-1}-e_{it-2})}{\sqrt{var(e_{it}-e_{it-1})}\sqrt{var(e_{it-1}-e_{it-2})}} = \frac{-\sigma_e^2}{2\sigma_e^2} = -0.5$
- 따라서 [xtserial]은 1차 차분 모형 오차항의 1계 자기상관계수가 -0.5인지를 검정한다
- findit xtserial
- xtserial gkor (inc mage mdur fkore cage cheal _I*), output
- [output] 옵션은 검정결과 만이 아닌 모형추정결과도 보여주라는 옵션이다
- 결과가 앞서 [vce] 옵션을 주고 추정했던 1차 차분 모형과 같다

```
Wooldridge test for autocorrelation in panel data
H0: no first order autocorrelation
    F(  1,     1380) =     81.690
           Prob > F =      0.0000
```

- 검정결과가 유의미하므로 영가설을 기각한다.
- 즉 e_{it} 에 자기상관이 존재한다.

41

5: 확률효과모형

- 기본 모형
- 모형 추정

42

기본 모형 1

- 여전히 기본 모형은 다음과 같다
- $y_{it} = \alpha + \beta x_{it} + u_i + e_{it}$
- 앞선 장의 고정효과 모형에서는 u_i를 추정해야할 모수(parameter)로 간주하였다
- 이에 반해 u_i를 확률변수(random variable)로 가정하는 것을 확률효과(random effects) 모형이라 한다
- 확률효과 모형에서 오차항들은 일반적으로 $u_i \sim N(0, \sigma_u^2)$과 $e_{it} \sim N(0, \sigma_e^2)$인 것으로 가정한다
- 모형을 다시 쓰면 $y_{it} = (\alpha + u_i) + \beta x_{it} + e_{it}$ 로 표현되며 $(\alpha + u_i)$는 확률변수로 간주되며 $E(\alpha + u_i) = \alpha$이다
- 달리 말해, α는 패널 개체별 상수항의 평균을 뜻하게 된다

43

기본 모형 2

- 이를 OLS로 추정하는 것은 1계 자기상관으로 인해 효율성 문제가 있다
  - $cov(u_i + e_{it}, u_i + e_{it-1}) = cov(u_i, u_i) + cov(u_i, e_{it-1}) + cov(e_{it}, u_i) + cov(e_{it}, e_{it-1}) = \sigma_u^2 \neq 0$
- OLS 추정을 위해서는 $cov(x_{it}, u_i) = 0$이어야 하는데 이것이 성립한다면 다음과 같은 모형을 추정하는 것이 일치추정량이면서 효율적인 추정량이라는 것이 알려져 있다
- $(y_{it} - \theta_i \bar{y}_i) = \alpha(1 - \theta_i) + \beta(x_{it} - \theta_i \bar{x}_i) + [u_i(1 - \theta_i) + (e_{it} - \theta_i \bar{e}_i)]$
- 여기에서 $\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i \sigma_u^2 + \sigma_e^2}}$
- 만약 $T_i = T$인 균형패널이면 $\theta_i = \theta$가 된다
- 먼저 $\hat{\sigma}_u^2$와 $\hat{\sigma}_e^2$를 추정한 다음, $\hat{\theta}_i$를 구한 후, 그것을 앞 식에 대입하여 OLS로 추정하는 단계를 거친다

44

기본 모형 3

- 그런데 고정효과 모형의 오차항 분산은$\sigma_{FE}^2 = var(u_i + e_{it}) = var(e_{it}) = \sigma_e^2$
- between 모형의 오차항 분산은 $\sigma_{BE}^2 = var(u_i + \bar{e}_i) = var(u_i) + var\left(\frac{\sum_{t=1}^{T_i} e_{it}}{T_i}\right) = \sigma_u^2 + \sigma_e^2/T_i$. 즉 $T_i\sigma_u^2 + \sigma_e^2 = T_i\sigma_{BE}^2$
- 따라서 θ_i는 다음과 같이 쓸 수 있다
- $\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i\sigma_u^2+\sigma_e^2}} = 1 - \sqrt{\frac{\sigma_{FE}^2}{T_i\sigma_{BE}^2}}$
- 확률효과 모형 추정량은 결국 between 모형의 추정량과 고정효과 모형의 추정량의 가중평균치라고 할 수 있다

45

기본 모형 4

- $\theta_i = 0$(즉 $\sigma_u^2 = 0$)이면 패널개체 별 이질성이 없다는 뜻이므로 합동 OLS로 추정하는 것과 같다
- 반면 $\theta_i = 1$(즉 $\sigma_e^2 = 0$)이면 오직 개체별 이질성만이 남기 때문에 u_i가 중요해지고 within 회귀모형과 같은 형태가 된다
- 확률효과 모형의 추정량은 between 모형 추정량보다 효율적이다
  - between 모형은 패널 별 평균만을 변수로 추정하는데 반해 확률효과 모형은 패널 간 정보와 패널 내 정보를 모두 사용한다
- $cov(x_{it}, u_i) = 0$이라는 가정이 성립하면
  - 고정효과 모형은 패널 개체 더미변수를 추정하기 때문에 자유도의 손실이 발생한다
  - 따라서 확률효과 모형의 추정량이 더 효율적이다
  - 또한 $\theta_i \neq 1$이면 시간에 따라 변하지 않는 설명변수(time invariant variable)에 대한 β 추정치도 얻을 수 있는 장점이 있다
- 하지만 $cov(x_{it}, u_i) = 0$이라는 가정이 성립하지 않는다면 확률효과 모형 추정량은 일치추정량이 되지 못한다

46

R

모형 추정 1

```
fit4 <- plm(gkor ~ factor(marst) +factor(mcont) + factor(medu) +
            inc + mage + mdur + factor(area) + fkore +
            factor(cgend) + cage + cheal,
          data=padt, model="random")
summary(fit4)
```

```
Unbalanced Panel: n = 1566, T = 1-9, N = 11547

Effects:
                  var std.dev share
idiosyncratic 0.4650  0.6819 0.688
individual    0.2113  0.4597 0.312
theta:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1708  0.5567  0.5567  0.5362  0.5567  0.5567

Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-2.79862 -0.44856  0.00324 -0.00006  0.45121  2.08258

Coefficients:
                    Estimate  Std. Error  z-value  Pr(>|z|)
(Intercept)     4.1075e+00  1.3647e-01  30.0976 < 2.2e-16 ***
factor(marst)1 -2.3298e-02  6.0358e-02  -0.3860 0.6994954
factor(marst)2 -3.6166e-02  6.4343e-02  -0.5621 0.5740590
factor(mcont)1 -1.1131e-02  5.8602e-02  -0.1899 0.8493598
factor(mcont)2  6.0115e-02  1.0191e-01   0.5899 0.5552589
factor(mcont)3 -4.8431e-02  5.8368e-02  -0.8297 0.4066803
factor(mcont)4 -1.0823e-03  5.8875e-02  -0.0184 0.9853332
factor(mcont)5 -2.4803e-03  8.6951e-02  -0.0285 0.9772435
factor(mcont)6 -3.3001e-02  7.7333e-02  -0.4267 0.6695704
factor(medu)1   5.6199e-02  4.6859e-02   1.1993 0.2303994
factor(medu)2   1.2247e-01  5.3710e-02   2.2803 0.0225903 *
factor(medu)3   2.0434e-01  5.6287e-02   3.6302 0.0002832 ***
inc             7.5562e-06  8.4928e-05   0.0890 0.9291049
mage            3.7382e-03  3.2872e-03   1.1372 0.2554561
mdur           -1.2380e-02  4.6147e-03  -2.6828 0.0073011 **
factor(area)1  -1.5369e-02  4.8366e-02  -0.3178 0.7506659
factor(area)2   3.0630e-02  5.1934e-02   0.5898 0.5553325
factor(area)3   3.0305e-02  5.0746e-02   0.5972 0.5503796
factor(area)4  -5.5100e-03  5.1461e-02  -0.1071 0.9147321
fkore           1.6782e-02  4.4537e-03   3.7680 0.0001645 ***
factor(cgend)1  1.5731e-01  2.7343e-02   5.7534 8.749e-09 ***
cage           -6.3425e-02  5.2862e-03 -11.9982 < 2.2e-16 ***
cheal          -9.8659e-02  1.1953e-02  -8.2537 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    6507
Residual Sum of Squares: 5372
R-Squared:      0.17443
```

47

Stata

모형 추정 2

```
xi: xtreg gkor i.marst i.mcont i.medu inc mage mdur i.area ///
fkore i.cgend cage cheal, re theta
```

- R과 Stata의 결과를 비교해 보면 약간 차이가 있는 것을 본다
- 예컨대 marst1의 계수가 R에서는 -0.0233이고 Stata에서는 -0.0237이다
- 이는 σ_{BE}^2를 추정하는 과정에서 R은 전체 관측사례들을 사용하지만 Stata에서는 between effect에서 사용한 사례들을 사용하기 때문인 것으로 보인다
- https://cran.rstudio.com/web/packages/plm/vignettes/B_plmFunction.html

```
Random-effects GLS regression                   Number of obs      =     11547
Group variable: pid                             Number of groups   =      1566

R-sq:  within  = 0.0789                         Obs per group: min =         1
       between = 0.0783                                        avg =       7.4
       overall = 0.0767                                        max =         9

                                                Wald chi2(22)      =    985.40
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

                            theta
------------------------------------------------------------------------------
       min       5%       median       95%       max
    0.1755    0.1755       0.5631    0.5631    0.5631
```

gkor	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Imarst_1	-.0237319	.0607149	-0.39	0.696	-.1427309	.0952671
_Imarst_2	-.0353876	.064676	-0.55	0.584	-.1621501	.091375
_Imcont_1	-.0103378	.0592824	-0.17	0.862	-.1265291	.1058535
_Imcont_2	.0603062	.1030982	0.58	0.559	-.1417626	.2623749
_Imcont_3	-.0484299	.0590396	-0.82	0.412	-.1641454	.0672856
_Imcont_4	-.0008353	.0595662	-0.01	0.989	-.1175828	.1159123
_Imcont_5	-.0023565	.0880007	-0.03	0.979	-.1748346	.1701216
_Imcont_6	-.0325418	.0782338	-0.42	0.677	-.1858773	.1207937
_Imedu_1	.0564041	.0473918	1.19	0.234	-.0364822	.1492903
_Imedu_2	.122902	.0543333	2.26	0.024	.0164108	.2293933
_Imedu_3	.2045974	.0569358	3.59	0.000	.0930053	.3161896
inc	5.44e-06	.0000851	0.06	0.949	-.0001614	.0001723
mage	.0037199	.0033263	1.12	0.263	-.0027995	.0102392
mdur	-.0123499	.004666	-2.65	0.008	-.0214952	-.0032047
_Iarea_1	-.0152319	.0488631	-0.31	0.755	-.1110018	.0805379
_Iarea_2	.0304612	.0525042	0.58	0.562	-.0724452	.1333676
_Iarea_3	.0303776	.051303	0.59	0.554	-.0701744	.1309297
_Iarea_4	-.0052602	.0520225	-0.10	0.919	-.1072223	.096702
fkore	.0166477	.0044567	3.74	0.000	.0079128	.0253827
_Icgend_1	.1572803	.0276712	5.68	0.000	.1030457	.2115149
cage	-.0634176	.005331	-11.90	0.000	-.0738662	-.0529691
cheal	-.0983411	.0119503	-8.23	0.000	-.1217632	-.074919
_cons	4.109067	.1376771	29.85	0.000	3.839224	4.378909
sigma_u	.46803302					
sigma_e	.68190044					
rho	.32023552	(fraction of variance due to u_i)				

48

6: Hausman 검정

- 하우스만 검정

49

하우스만 검정 1

- 패널 모형은 다음과 같이 썼다
- $y_{it} = (\alpha + u_i) + \beta x_{it} + e_{it}$
- 지금까지 배운 것처럼 고정효과 모형에서는 $\alpha + u_i$를 패널 개체별로 고정되어 있는 모수로 해석한다
- 이에 반해 확률효과 모형에서는 $\alpha + u_i$를 확률분포를 따르는 확률변수로 해석한다. 즉 $(\alpha + u_i) \sim N(\alpha, \sigma_u^2)$
- 고정효과와 확률효과 중 어떤 모형을 선택할 것인지는 첫째로 u_i에 대한 해석에 달려있다
  - 패널 개체들이 모집단에서 무작위로 추출된 표본의 개념이라면 오차항 u_i는 확률분포를 따른다고 가정할 수 있다
  - 하지만 패널 개체들이 모집단에서 추출된 표본이 아니라 특정 모집단 그 자체라면 u_i는 확률분포를 따른다고 할 수 없다
  - 다문화청소년패널 자료를 분석할 때 모집단에서 추출하였기 때문에 확률효과를 사용하는 것이 바람직해 보인다
  - 이에 반해 OECD국가에 대한 분석이나 미국의 50개 주 패널에 대한 자료는 그 자체로 모집단을 형성하기 때문에 고정효과로 보는 것이 적절하다

50

하우스만 검정 2

- 통계적 측면에서 보았을 때, $cov(x_{it}, u_i) = 0$이면
  - 고정효과 추정량과 확률효과 추정량이 모두 일치추정량이다. 따라서 두 모형의 추정량이 유사할 것이다
- $cov(x_{it}, u_i) \neq 0$이면
  - 고정효과 추정량은 여전히 일치추정량이지만 확률효과 추정량은 일치추정량이 아니다. 따라서 두 추정량 사이에 차이가 존재할 것이다
- 하우스만(Hausman) 검정은 이러한 특성을 이용한다.
  - 즉 $H_0: cov(x_{it}, u_i) = 0$ 하에 두 추정량은 유사하지만 대안가설 하에 두 추정량은 차이가 있다
  - $H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})' [var(\hat{\beta}_{FE}) - var(\hat{\beta}_{RE})]^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{RE}) \sim \chi^2_{K-1}$를 이용한다
  - 하우스만 검정에 있어 $var(\hat{\beta}_{FE}) - var(\hat{\beta}_{RE})$ 행렬이 양정행렬(positive definite)이 되어야 하지만 그렇지 않은 경우에도 큰 문제는 없는 것으로 알려져 있다
  - 만약 양정행렬이 되지 않으면 검정통계치 χ^2_{K-1}가 음수가 될 수 없음에도 불구하고 현실적으로 음수 값을 가질 수 있다. 이 경우 sigmamore 혹은 sigmaless 옵션을 사용해 본다

51

R

하우스만 검정 3

```
phtest(gkor ~ factor(marst) + factor(medu) +
          inc + mage + mdur + factor(area) + fkore +
          cheal,
        data=padt, model=c("within","random"))
```

```
        Hausman Test

data:  gkor ~ factor(marst) + factor(medu) + inc + mage + mdur + factor(area) +  ...
chisq = 174.53, df = 14, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent
```

52

Stata

하우스만 검정 3

- qui xi: xtreg gkor i.marst i.medu inc mage ///
- mdur i.area fkore cheal, fe
- estimates store FE

- qui xi: xtreg gkor i.marst i.medu inc mage ///
- mdur i.area fkore cheal, re
- estimates store RE

- hausman FE RE /* 반드시 FE를 먼저 적는다 */

	Coefficients (b) FE	(B) RE	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
_Imarst_1	-.0489212	-.0813163	.0323951	.0697873
_Imarst_2	.0257583	-.0884496	.1142079	.0634601
_Imedu_1	.1447087	.1119016	.0328071	.2559078
_Imedu_2	.5602797	.1972936	.3629861	.7119933
_Imedu_3	.8459747	.2446433	.6013314	.545591
inc	-.0001231	-.0001599	.0000368	.0000589
mage	-.0451362	-.0053129	-.0398234	.0268303
mdur	-.0261167	-.0488634	.0227467	.0269534
_Iarea_1	-.0647642	-.0374431	-.0273211	.1406978
_Iarea_2	-.2074424	.0100047	-.2174471	.2284356
_Iarea_3	.1326215	.0089924	.1236291	.2731901
_Iarea_4	.0786572	-.0230089	.1016661	.2198155
fkore	.0105064	.0222142	-.0117077	.0024837
cheal	-.0853496	-.1031524	.0178028	.0042341

b = consistent under Ho and Ha; obtained from xtreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

chi2(13) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= 140.06
Prob>chi2 = 0.0000

해석?

53

7: Multilevel model

- Multilevel model
- 모형 추정

54

Multilevel model 1

- 종단 자료는 한 개인에 대해 여러 시점의 응답이 있으므로 다양한 시점의 응답들은 개인들에 놓여 있다(nested)고 볼 수 있다
- 이런 점에서 특정 시기의 응답은 Level 1으로 볼 수 있고 개인은 Level 2로 볼 수 있기 때문에 다층모형을 이용한 분석이 가능하다
- 설명 변수가 하나인 기본적인 다층모형은 다음과 같이 나타낸다
- $y_{it} = \beta_{0i} + \beta_1 x_{1it} + e_{it}$
- $\beta_{0i} = \gamma_{00} + u_{0i}$
- 위 모형에서 종속 변수에 주는 설명 변수의 계수는 개인별로 변하지 않지만 절편(intercept)이 개인별로 변하는 모형이며 이를 random intercept 모형이라고 한다
- 앞서 배운 확률효과 모형과 같지만 다층모형에서는 추정 방법이 달라 계수나 통계적 추론에 있어 같지 않다

55

R

모형 추정 1

```
install.packages("lme4")
library(lme4)
summary(md <- lmer(gkor ~ factor(marst) +factor(mcont) + factor(medu) +
          inc + mage + mdur + factor(area) + fkore +
          factor(cgend) + cage + cheal + (1 | pid),
        data=adt))
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 pid      (Intercept) 0.2147   0.4634
 Residual             0.4655   0.6822
Number of obs: 11547, groups:  pid, 1566

Fixed effects:
                Estimate Std. Error t value
(Intercept)     4.108e+00  1.370e-01  29.993
factor(marst)1 -2.348e-02  6.051e-02  -0.388
factor(marst)2 -3.584e-02  6.448e-02  -0.556
factor(mcont)1 -1.080e-02  5.888e-02  -0.183
factor(mcont)2  6.019e-02  1.024e-01   0.588
factor(mcont)3 -4.843e-02  5.865e-02  -0.826
factor(mcont)4 -9.791e-04  5.916e-02  -0.017
factor(mcont)5 -2.429e-03  8.739e-02  -0.028
factor(mcont)6 -3.281e-02  7.771e-02  -0.422
factor(medu)1   5.628e-02  4.708e-02   1.196
factor(medu)2   1.227e-01  5.397e-02   2.273
factor(medu)3   2.044e-01  5.656e-02   3.615
inc             6.673e-06  8.501e-05   0.078
mage            3.731e-03  3.303e-03   1.129
mdur           -1.237e-02  4.636e-03  -2.668
factor(area)1  -1.531e-02  4.857e-02  -0.315
factor(area)2   3.056e-02  5.217e-02   0.586
factor(area)3   3.034e-02  5.098e-02   0.595
factor(area)4  -5.406e-03  5.169e-02  -0.105
fkore           1.673e-02  4.455e-03   3.754
factor(cgend)1  1.573e-01  2.748e-02   5.724
cage           -6.342e-02  5.305e-03 -11.956
cheal          -9.853e-02  1.195e-02  -8.243
```

56

Stata

모형 추정 2

- xi: xtmixed gkor i.marst i.mcont i.medu inc mage mdur i.area ///
- fkore i.cgend cage cheal || pid :, reml
- R의 추정방법은 Restricted maximum likelihood 방법이기 때문에 Stata에서도 REML 옵션을 주었고 결과는 같아 보인다
- R에서는 e_{it}와 u_{0i}의 분산을 보여주지만 Stata에서는 표준편차를 보여준다는 점이 다른 것 같다

```
                                                Wald chi2(22)     =     985.33
Log restricted-likelihood = -13139.805          Prob > chi2       =     0.0000
```

gkor	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Imarst_1	-.0234791	.0605069	-0.39	0.698	-.1420704	.0951121
_Imarst_2	-.0358422	.0644822	-0.56	0.578	-.162225	.0905405
_Imcont_1	-.0107996	.058884	-0.18	0.854	-.1262101	.1046109
_Imcont_2	.0601947	.1024011	0.59	0.557	-.1405078	.2608973
_Imcont_3	-.0484305	.0586464	-0.83	0.409	-.1633754	.0665143
_Imcont_4	-.0009791	.0591614	-0.02	0.987	-.1169334	.1149752
_Imcont_5	-.0024286	.0873865	-0.03	0.978	-.1737029	.1688457
_Imcont_6	-.0328091	.0777064	-0.42	0.673	-.1851109	.1194927
_Imedu_1	.0562847	.0470799	1.20	0.232	-.0359903	.1485596
_Imedu_2	.1226523	.0539683	2.27	0.023	.0168765	.2284281
_Imedu_3	.2044442	.0565562	3.61	0.000	.093596	.3152924
inc	6.67e-06	.000085	0.08	0.937	-.00016	.0001733
mage	.0037305	.0033034	1.13	0.259	-.002744	.0102051
mdur	-.0123675	.004636	-2.67	0.008	-.0214538	-.0032812
_Iarea_1	-.0153112	.0485721	-0.32	0.753	-.1105109	.0798884
_Iarea_2	.0305607	.0521708	0.59	0.558	-.0716922	.1328135
_Iarea_3	.0303357	.0509773	0.60	0.552	-.0695779	.1302492
_Iarea_4	-.005406	.0516941	-0.10	0.917	-.1067246	.0959125
fkore	.0167256	.004455	3.75	0.000	.007994	.0254571
_Icgend_1	.1572983	.0274788	5.72	0.000	.1034409	.2111558
cage	-.0634218	.0053047	-11.96	0.000	-.0738189	-.0530247
cheal	-.0985256	.011952	-8.24	0.000	-.1219511	-.0751
_cons	4.108132	.1369708	29.99	0.000	3.839674	4.37659

Random-effects Parameters		Estimate	Std. Err.	[95% Conf.	Interval]
pid: Identity					
	sd(_cons)	.4633718	.0112422	.4418532	.4859383
	sd(Residual)	.6822489	.0048205	.672866	.6917627

LR test vs. linear regression: chibar2(01) = 2093.81 Prob >= chibar2 = 0.0000

57

Multilevel model 2

- 설명 변수들의 계수 또한 개인에 따라 변할 수 있다
- $y_{it} = \beta_{0i} + \beta_{1i}x_{1it} + e_{it}$
- $\beta_{0i} = \gamma_{00} + u_{0i}$
- $\beta_{1i} = \gamma_{10} + u_{1i}$
- 위 모형은 설명 변수의 계수(coefficient)이 개인별로 변하는 모형이라 random coefficient 모형이라고 한다

58

R

모형 추정 3

```
summary(md <- lmer(gkor ~ factor(marst) +factor(mcont) + factor(medu) +
                   inc + mage + mdur + factor(area) + fkore +
                   factor(cgend) + cage + cheal +
                    (1 + fkore + cage| pid),
                  data=adt))
```

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 pid      (Intercept) 1.238737 1.11299
          fkore       0.001665 0.04081  -0.17
          cage        0.006324 0.07952  -0.86 -0.16
 Residual             0.418062 0.64658
Number of obs: 11547, groups:  pid, 1566

Fixed effects:
                    Estimate Std. Error t value
(Intercept)      4.126e+00  1.385e-01  29.782
factor(marst)1 -3.001e-02  6.216e-02  -0.483
factor(marst)2 -2.936e-02  6.820e-02  -0.431
factor(mcont)1 -9.807e-04  5.982e-02  -0.016
factor(mcont)2  4.516e-02  1.007e-01   0.448
factor(mcont)3 -5.110e-02  5.856e-02  -0.873
factor(mcont)4 -1.807e-03  5.936e-02  -0.030
factor(mcont)5  5.842e-03  8.611e-02   0.068
factor(mcont)6 -2.214e-02  7.724e-02  -0.287
factor(medu)1   5.779e-02  4.740e-02   1.219
factor(medu)2   1.244e-01  5.401e-02   2.304
factor(medu)3   1.886e-01  5.666e-02   3.330
inc             4.429e-05  8.722e-05   0.508
mage            3.356e-03  3.267e-03   1.027
mdur           -1.322e-02  4.616e-03  -2.864
factor(area)1  -1.822e-02  4.876e-02  -0.374
factor(area)2   2.439e-02  5.217e-02   0.467
factor(area)3   2.399e-02  5.095e-02   0.471
factor(area)4   5.367e-03  5.174e-02   0.104
fkore           1.621e-02  4.617e-03   3.512
factor(cgend)1  1.604e-01  2.730e-02   5.877
cage           -6.301e-02  5.668e-03 -11.117
cheal          -9.768e-02  1.192e-02  -8.193
```

59

Stata

모형 추정 4

- xi: xtmixed gkor i.marst i.mcont i.medu inc mage mdur i.area ///
- fkore i.cgend cage cheal || pid : fkore cage, reml cov(unstr)

gkor	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Imarst_1	-.0300017	.0621617	-0.48	0.629	-.1518363	.0918329
_Imarst_2	-.0293604	.0682045	-0.43	0.667	-.1630387	.1043179
_Imcont_1	-.0010005	.0598158	-0.02	0.987	-.1182373	.1162363
_Imcont_2	.0451395	.1007389	0.45	0.654	-.1523053	.2425842
_Imcont_3	-.0510901	.0585557	-0.87	0.383	-.1658572	.063677
_Imcont_4	-.0018496	.0593605	-0.03	0.975	-.1181941	.114495
_Imcont_5	.0057235	.0861112	0.07	0.947	-.1630515	.1744984
_Imcont_6	-.0221553	.0772375	-0.29	0.774	-.173538	.1292274
_Imedu_1	.0577872	.047397	1.22	0.223	-.0351093	.1506837
_Imedu_2	.1244167	.0540106	2.30	0.021	.0185578	.2302756
_Imedu_3	.1886262	.0566563	3.33	0.001	.0775818	.2996706
inc	.0000443	.0000872	0.51	0.611	-.0001266	.0002153
mage	.003361	.0032669	1.03	0.304	-.0030421	.0097641
mdur	-.0132264	.004616	-2.87	0.004	-.0222736	-.0041793
_Iarea_1	-.0182211	.0487564	-0.37	0.709	-.1137819	.0773397
_Iarea_2	.0244314	.0521718	0.47	0.640	-.0778234	.1266862
_Iarea_3	.0240707	.0509451	0.47	0.637	-.07578	.1239214
_Iarea_4	.0053487	.0517387	0.10	0.918	-.0960572	.1067546
fkore	.0162118	.0046195	3.51	0.000	.0071578	.0252657
_Icgend_1	.1604223	.0272983	5.88	0.000	.1069186	.2139261
cage	-.0630106	.005669	-11.12	0.000	-.0741216	-.0518997
cheal	-.0976778	.0119219	-8.19	0.000	-.1210444	-.0743113
_cons	4.126141	.1385825	29.77	0.000	3.854524	4.397757

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
pid: Unstructured				
sd(fkore)	.0411233	.011209	.0241032	.0701618
sd(cage)	.0795885	.003574	.0728831	.0869109
sd(_cons)	1.117898	.1006634	.9370306	1.333676
corr(fkore,cage)	-.1584081	.1606319	-.448395	.1617497
corr(fkore,_cons)	-.1785306	.2249231	-.5620462	.2681693
corr(cage,_cons)	-.8634802	.0528732	-.9371622	-.7160641
sd(Residual)	.6465081	.0050363	.6367121	.6564548

LR test vs. linear regression: chi2(6) = 2355.35 Prob > chi2 = 0.0000

60

Thank You

- 발표를 들어 주셔서 감사합니다
- 자료를 수집하고 배포하느라 고생하시는 다문화청소년패널조사 관계자님들께 감사합니다
- 질문이나 자료에 오류가 있다면 다음 이메일로 알려주세요
- sochyunsik@khu.ac.kr

61

MEMO

MEMO

MEMO