계량분석
Examining Data

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

September 27, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. 통계 데이터(Statistical Data)
2. 데이터 아카이브(Data Archive)
3. 일반사회조사(General Social Survey)
4. 공공 데이터(Public Data)
5. 패널 데이터(Panel Data)
6. Takeaways

통계 데이터(Statistical Data)

통계 데이터(Statistical Data)

통계 자료(statistical data)는 직접 수집한 1차 자료와 남이 수집한 2차 자료(secondary data)로 구분할 수 있다.

- 편의표본(convenient sample)에 근거하여 자기가 직접 1차 자료를 수집한 뒤, 이를 분석하여 논문을 쓰는 경우도 제법 많다.
- 증거 자료의 체계적인 수집과 정리, 관리도 하나의 전문적인 경험과학자의 역할이라고 언급하였다. 이에 요구되는 전문적인 능력, 자본, 시간 없이 수집된 1차 자료는 현실적으로 기준 미달인 경우가 많다.
- 이러한 데이터에 의존한 발견과 결론은 일반화(generalization) 될 수 없고 그만큼 가치가 떨어진다.

그러나 1차 자료의 수집(과 공개) 자체가 훌륭한 연구(의 일부)로 인정받는 경우도 있다.

- 내용분석(content analysis)을 수행하는 경우에는 대부분 직접 1차 자료를 수집해야 한다.
- 설문조사(survey)의 경우에도 2차 자료로는 존재하지 않는 숨겨진 모집단(hidden population)을 연구한다면 직접 1차 자료를 수집해야 한다.
- 다만 데이터를 직접 수집하다보면 석박사 과정의 기간이 (아무리 적게 잡아도) 최소한 1년은 늘어난다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

통계 데이터(Statistical Data)

통계 데이터는 수집 절차에 따라 다음과 같이 나뉜다.

- 실험(experimental) 데이터
- 비실험(non-experimental; observational) 데이터
- 모의실험(simulation) 데이터

많은 사회과학 연구는 주로 비실험 데이터를 다루어왔다.

- 그러나 이것은 실험 데이터와 모의실험 데이터를 고려할 필요가 없다는 의미는 아니다.
- 다만 모의실험 데이터는 수리사회학과 수리통계학의 영역이므로 기회가 닿는다면 다른 수업에서 다루기로 한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

통계 데이터(Statistical Data)

실험 데이터와 비실험 데이터는 다르다.

- 실험(experiment)은 상대적으로 소수의 참가자들(subjects)을 선발한다. 비실험 데이터는 상대적으로 다수를 담고 있다.
- 실험 참가자들은 무작위(random)로 처방집단(treatment group)과 통제집단(control group)으로 나뉜다. 비실험 데이터에서도 처방집단과 통제집단이 나뉘지만 여기에는 무작위라는 보장이 없다.
- 실험 결과, 실험집단과 통제집단 사이에서 나타나는 차이는 처방의 인과적 효과(treatment effect)로 귀속된다. 비실험 데이터의 경우 그렇지 않다.
- 실험의 결과(results)가 실험 참가자들 이외의 사람들에게 일반화(generalization)될 수 있는가는 다소 애매하다. 그렇기 때문에 메타분석(meta analysis)이 필요하다. 무작위 표본(random sample)을 사용한 비실험 데이터는 적어도 상관관계의 일반화는 잘 된다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

통계 데이터(Statistical Data)

실험 데이터에서 무작위화(randomization)와 비실험 데이터에서 무작위 표본(random sample)은 분명히 다르다.

- 실험 데이터의 randomization는 어떤 subject가 treatment group과 control group 중 "오로지 우연에 의해서" 어느 쪽에 속할지 결정된다는 의미이다.
- Randomization 덕택에 실험 연구(experiment)에서 처방(treatment)과 효과(effect) 사이에 인과관계(causal relations)가 인정된다.
- 비실험 데이터에서 무작위 표본(random sample)이란 모집단(population)의 구성원이 "오로지 우연에 의해' 서' 선발된 표본(sample)을 뜻한다. 단순 무작위 표본(simple random sample)에서 population의 구성원이 표본으로 뽑힐 확률은 동일하다.
- Random sample 덕택에 표본 안에서의 분석 결과는 모집단 전체로 일반화 될 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

통계 데이터(Statistical Data)

통계 데이터는 수집 방식에 따라 다음과 같이 나뉜다.

- 횡단면(cross-sectional) 데이터
- 시계열(time-series) 데이터
- 패널(panel) 데이터
- 반복된 횡단면(repeated cross-sectional) 데이터

많은 사회과학 연구는 주로 횡단면 데이터를 다루어왔다.

- 그러나 이것은 시계열, 반복된 횡단면, 그리고 패널 데이터를 고려할 필요가 없다는 의미는 아니다.
- 다만 그것들은 난이도가 높으므로 기회가 닿는다면 다른 수업에서 다루기로 한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

통계 데이터(Statistical Data)

횡단면(cross-sectional) 데이터는 "주어진 시간대"에 여러 분석대상(subjects)에 관해 조사한 자료이다.

- 분석대상은 사람, 고양이, 암세포 밀도, 단어의 수 등 연구 목적에 따라 다양하다.
- 이 데이터가 수집되는 동안 시간 변화에 따른 차이는 고려되지 않는다.

시계열(time-series) 데이터는 "시간이 경과함에 따라" 하나의 분석대상의 변화하는 특성을 조사한 자료이다.

- 대표적인 예로 일별 주가지수나 연간 강수량 등을 생각해 볼 수 있다.
- 이 데이터에서 핵심은 시간이 흐르는 동안의 변화가 계속 기록된다는 점이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

통계 데이터(Statistical Data)

패널(panel) 데이터는 여러 분석대상에 관해 “시간이 경과함에 따라” 추적하여 반복 조사한 자료이다.

- 여기서 중요한 것은 추적하였다는 점이다. 다시 말해, “여러 분석대상”은 시간 경과에 따라 재방문하여 재조사되었다.
- 패널 데이터에서 “분석대상”이 반드시 사람이여야 하는 것은 아니다. 동네(town)나 국가(country)와 같은 장소일 수도 있고 사건(event)일 수도 있다.

반복된 횡단면(repeated cross-sectional) 데이터는 여러 분석대상에 관해 “시간이 경과함에 따라” 추적하지는 않고 반복 조사한 자료이다.

- 추적이 이루어지지 않았으므로, “여러 분석대상”은 조사 시점에 따라 매번 다르다 (우연히 같은 사람이 걸렸을 수는 있다).

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

통계 데이터(Statistical Data)

통계 데이터는 여러 포멧(formats)을 취할 수 있다.

- 분석에 투입되기 이전의 순간 데이터는 반드시 숫자의 나열(수열) 형태를 갖추어야 한다.
- 하지만 아직 수열로 전환되기 이전의 자료도 일종의 데이터로 볼 수 있다. 가령 사진, 그림, 비디오, 오디오, 문서/일기/신문, 파일, 움직임에 관한 기록, 질적 자료 (qualitative data), 심지어 통계 데이터에 관한 설명 그 자체도 데이터가 될 수 있다.
- 우리는 pre-processing 이라는 절차를 통해 위와 같은 데이터를 수열의 꼴로 바꿀 수 있다. 물론 pre-processing 절차 그 자체도 배워야 한다.
- 데이터의 정의는 오로지 자신의 상상력의 한계에 의해서만 제약받는다.

이 수업에서는 pre-processing을 생략하기 위해 수열의 꼴로 주어진 데이터를 바로 분석한다.

이렇게 분석에 ready-to-go 상태에 있는 데이터를 데이터셋(dataset) 이라고도 부른다. 현실적으로 연구자들 사이에서 data와 dataset을 그리 엄격하게 구분하지는 않는다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터 아카이브(Data Archive)

데이터 아카이브(Data Archive)

경험적 사회과학이 성장하고 가정용 컴퓨터(PC)가 보급된 1990년대 이후의 시점에서 데이터 아카이브(data archive)가 탄생했다.

- 자신이 수집한 자료를 한 번 쓰고 버리기보다 남들도 연구와 교육을 위해 쓸 수 있도록 배려한다.
- 경험과학의 성격상 자료 수집 자체에 전문화된 사람들이 다양한 주제로 데이터를 수집하여 공개하거나 판매하였다.
- 데이터 아카이브의 출현은 계량적 사회과학(quantitative social sciences)의 출현과 발달에 가장 중요한 인프라였다.
- 어떤 조직이 데이터 아카이브를 운영하고 통제하는 것은 말하자면 그 조직이 계량적 사회과학 계에서 기축통화를 가졌다는 말과 같다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터 아카이브(Data Archive)

종합적인 주제를 모두 커버하는 데이터 아카이브와 특수한 주제만을 커버하는 데이터 아카이브로 나뉜다.

- 전세계에서 명실상부 가장 대표적인 종합형 데이터 아카이브는 Inter-university Consortium for Political and Social Research로 이른바 ICPSR (https://www.icpsr.umich.edu)라고 불리운다.
- Pew Research Center (https://www.pewresearch.org)
- Roper Center (https://ropercenter.cornell.edu)
- Harvard Dataverse (https://dataverse.harvard.edu)
- 우리나라에서 대표적인 종합형 데이터 아카이브는 2021년 현 시점에서 아마도 한국사회과학자료원(https://kossda.snu.ac.kr)과 한국사회과학데이터센터 (https://www.ksdc.re.kr)인 것 같다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터 아카이브(Data Archive)

- 특수주제형 데이터 아카이브는 수가 무척 많고 여기저기 흩어져 있어서 자기 전공 분야만 잘 아는 경우가 많다.
- 경제 데이터의 경우 National Bureau of Economic Research (NBER)가 유명하다 (https://www.nber.org/research/data).
- 종교 데이터의 경우 The Association of Religion Data Archives (ARDA)가 유명하다(https://thearda.com).
- 경영/금융 데이터의 경우 Wharton Research Data Services (WRDS)가 유명하지만 기관 라이선스가 필요하다.
- 인구 데이터의 경우 Social Explorer가 제법 유명하고 편리하지만 기관 라이선스가 필요하다.

시간을 많이 들여서 관심에 부합하는 데이터 아카이브를 발굴하고 자주 살펴보는 습관이 필요하다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터 아카이브(Data Archive)

여기서는 여러 데이터 아카이브 가운데 한국사회과학데이터센터(KSDC)를 방문해 보기로 하자.

- 현재 충북대학교 도서관은 홈페이지 관리 부실로 KSDC 링크가 깨진 상태를 내버려두고 있다.
- 교내라면 https://ksdcdb.kr를 입력해 직접 들어갈 수 있다. 교외의 경우 교외 접속을 하고 직접 입력해 따로 들어가야 한다.

여기서 "[2699] 마스크 및 사회적 거리두기에 대한 대국민 인식조사"를 다운받자. 원자료 뿐 아니라 코드북(codebook)이나 설문지(questionnaire)도 함께 다운받아야 한다(**[Stata 코드]** 참고).

- 먼저 설문지를 쭉 살펴보면서 어떤 문항들이 있는지 살펴보자. 무슨 변수들이 독립변수(independent variable)로, 또 종속변수(dependent variable)로 어울릴지 상상해보자.
- 무슨 이론이나 가설이나 두 변수 사이의 관계에 대해 어떤 식으로 시사하나?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터 아카이브(Data Archive)

- Stata가 구버전이라서 **import spss** 명령어를 사용할 수 없으면 SPSS를 기동시켜 "새로 저장하기" 기능을 활용해 Stata 파일로 저장하자.
- "문5. 귀하는 아래의 대상 및 목적으로 마스크를 수출하는 것에 대해 어떻게 생각하십니까?"를 살펴보자. 이 안의 다섯개 문항은 마스크 수출 반대라는 태도/의견의 여러 차원들(dimensions)을 측정하고 있다. 각각의 변수에 대해 빈도분포표(frequency distribution table)를 살펴보고 해석해보자.
- 마스크 수출 반대라는 개념의 차원을 이론적으로 제대로 측정하고 있는가는 지금 당장 고민하지 말자. 그저 연습삼아 이들 변수를 모두 더해 하나의 변수를 만들어보고 그것의 빈도분포표를 살펴보자. 이 새로운 변수에서 큰 값을 무엇을 의미하나?
- 새로운 변수(**oppose**)의 히스토그램(histogram)을 그려보자.
- **oppose**과 또다른 변수인 응답자의 연령 간의 연관성을 산포도(scatterplot)로 간단히 살펴보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터 아카이브(Data Archive)

이 데이터에서 “문4. 귀하는 마스크 5부제에 대해 어떻게 평가하십니까?”를 보자(**[Stata 코드]** 참고).

- 개별 변수들의 frequency distribution table을 그려보자. 해석해보자.
- 이들 변수들을 역코딩(reverse coding)하여 새로운 변수를 만들어보자. 이제 이 변수들을 해석해보자.
- 연습삼아 역코딩한 변수들을 모두 더하여 하나의 변수로 만들어보자. 이 값이 크다는 것은 무엇을 의미하는지 해석해보자.
- 이 변수의 histogram을 그려보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일반사회조사(General Social Survey)

일반사회조사(General Social Survey)

데이터 아카이브 안에서 이른바 "일반사회조사(General Social Survey)"를 찾을 수 있다.

- 해당 국가의 거주민들이 정치 · 사회 · 경제 · 과학기술 등 제분야에 걸친 가치(values), 여론(opinions), 태도(attitudes), 행태(behaviors)에 대해 조사한 자료이다.
- 시민 자유(civil liberties), 범죄/폭력, 그룹간 관용(intergroup tolerance), 도덕/윤리 판단, 국가재정 지출, 심리적 안녕(psychological well-being), 사회적 계층 이동(social mobility), 스트레스와 트라우마 등 폭넓은 토픽을 물어본다. 물론 나이, 성별, 인종 등 기본적인 인구학적 변수(demographic variables)를 포함한다.
- 특정 토픽을 수 년에 한 번씩 돌리기도 한다.
- 일반적으로 특정 국가 내 모든 거주민들에 대해 대표성을 확보하고 있다(nationally representative). 국가별로 상이하다(미국은 GSS, 한국은 KGSS, 중국은 CGSS, 일본은 JGSS, 독일은 GGSS 등등).
- 유사한 질문을 수 년에 걸쳐 반복적으로 질문하여 사회변동을 파악하기에 유용하다. 다만 같은 사람을 추적하지 않으므로 꼭 패널(panel)인 것은 아니다. 다만 미국의 GSS는 패널을 운용하고 있다.
- 데이터 수집 및 관리를 전공한 수많은 사회학자들이 프로젝트를 운영하기 때문에 일반적으로 대단히 퀄리티가 높다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일반사회조사(General Social Survey)

일반사회조사 데이터들을 비교하여 국가간 비교(cross-national comparison)를 수행할 수도 있다.

- 이 목적으로 설계·수집된 대표적인 데이터는 Ronald Inglehart가 주도한 World Values Survey (WVS)다(https://www.worldvaluessurvey.org).
- 유럽 국가들 사이에서 설계·수집된 데이터로는 European Values Survey (https://europeanvaluesstudy.eu)가 있다.
- 몇몇 국가의 연구자들이 자국에서 GSS를 운영하면서 같은 모듈(module)을 같은 해에 함께 질문하여 일부러 비교가능하도록 설계한다. 이 데이터를 따로 뽑아 International Social Survey Programme (ISSP)을 구축했다 (http://w.issp.org/menu-top/home).
- 많은 유럽 국가들은 ESS (European Social Survey)에 함께 참여한다 (https://www.europeansocialsurvey.org).
- 동아시아의 국가들 간에 운영되는 East-Asian Social Survey (EASS)도 있다 (https://www.eassda.org).
- (필요에 따라) 한국인과 미국인의 사회적 가치를 비교하기 위해 KGSS와 GSS를 함께 분석할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일반사회조사(General Social Survey)

오늘날 비교사회학(comparative sociology)은 다소 침체되었지만 폭넓은 데이터에 힘입어 다양한 주제를 탐구하며 여전히 명맥을 유지하고 있다.

- 국민정체성(national identity), 민족주의(nationalism), 그리고 이주민에 대한 태도(attitudes toward migrants)
- 국가기구에 대한 신뢰(confidence), 대중 일반에 대한 신뢰(trust)
- 노동조합 가입률 및 조직관련 행동
- 결혼과 가족, 여성의 사회적 지위, 아동 양육 등에 관련한 태도 및 행태, 가치관
- 더 많은 시민적 자유(civil liberties)와 통치가능성(governability) 사이에서의 믿음
- 정치 및 사회 참여, 신사회운동(new social movements) 가치관
- 교육, 보건, 의료와 관련된 태도 및 행동

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일반사회조사(General Social Survey)

일반사회조사 가운데에서 세계가치관조사(WVS)에서 수집된 가장 최근의 데이터를 시험삼아 다운로드 받자.

- 링크는 https://www.worldvaluessurvey.org
- Stata 데이터 파일 뿐만 아니라 설문지(Questionnaire)도 다운받을 것.

설문지를 먼저 꼼꼼히 들여다 보자.

- 자신의 이론 또는 가설에 따른 키워드를 검색하여 독립변수와 종속변수 등을 찾아낸다.
- 수많은 데이터셋에서 조사하는 변수들을 미리 머리 속에 잘 정리해두고 있다가, 이론 또는 가설이 생겨났을 때 "아! 이 가설이라면 WVS가 적절하겠구나!" 하고 깨닫는 것이 보다 이상적이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일반사회조사(General Social Survey)

우리는 특히 탈물질주의적 가치(post-materialist values)에 관해 살펴보기로 한다(**[Stata 코드]** 참고). 다음의 언명(statement) 가운데 가장 중요한 두 가지가 무엇인가를 고르게 하여 분류한다(1=물질주의적; 2=혼합적; 3=탈물질주의적).

1. Maintain order in the country
2. Give people more to say in important government decisions
3. Fight raising prices
4. Protect freedom of speech

잉글하트, 로널드. 1983.『조용한 혁명』. 종로서적.

잉글하트, 로널드 · 크리스찬 웰젤. 2011.『민주주의는 어떻게 오는가: 근대화, 문화적 이동, 가치관의 변화로 읽는 민주주의의 발전 지도』. 김영사.

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

공공 데이터(Public Data)

공공 데이터(Public Data)

사회학 연구에서 정부의 공식 통계(official statistics)의 쓸모는 다소 미묘한 입장에 놓여왔다.

- Durkheim의 〈자살론〉은 공식 통계를 활용한 가장 뛰어난 사회학 고전 연구다. 하지만 공식 통계에 의존했다는 이유로 비판받기도 했다.
- 사회학 연구에서 공식 통계의 사용에 관한 가장 근본적인 비판 가운데 하나는 민속방법론(ethnomethodology)에 의해 제기되었다: "우리는 사회 현상을 통계적으로 분석하는가? 아니면 공무원의 통계 작성 행위를 분석하는가?"
- 설령 통계 자료의 중립성을 받아들이더라도 공식 통계가 대부분 집계 자료(aggregate data)라는 점에서 유용성이 다소 제한적이다. 어떤 경우에는 집계 자료로도 충분하지만 원자료(raw data)가 필요한 경우가 많다.
- 원자료는 프라이버시나 저작권 등의 이유로 인해 공공 데이터로서는 공개되지 않지만 연구 목적에 따라서는 구입 또는 무료로 확보할 수 있는 경우도 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

공공 데이터(Public Data)

근래에 들어 빅데이터(big data)와 비정형(unstructured) 데이터(unstructured)의 형식으로 제공하는 자료의 양과 범위가 점점 넓어지고 있다.

- 이른바 4차 산업혁명의 한 인프라로서 공공 데이터의 가치가 재발견되면서 연구 기회도 늘어났다. 각종 경진대회가 열리기도 한다. 예컨대 2021년 제9회 문화공공데이터 활용 경진대회(https://www.culture.go.kr/contest/main.do).
- 비정형인 데이터가 일반화되면서 좀 더 새롭고 창의적인 접근이 요구되는 경우가 많아졌다. 연구자의 pre-processing 스킬도 중요해졌다.
- 다운로드가 아니라 API (Application Programming Interface)의 형식을 취하는 경우가 늘어났다. 이런 경우에는 데이터에 접근하기 위해서라도 프로그래밍(주로 R이나 Python)을 배워야 한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

공공 데이터(Public Data)

여기서는 세계은행(World Bank)에서 제공하는 데이터를 살펴보자.

- https://data.worldbank.org/
- World Development Indicators 메뉴를 선택하자. 모든 Country, 모든 Series, 2019년을 체크해서 국가 단위의 경제 통계를 csv 파일의 형식으로 다운로드받자.
- 다운로드받은 데이터를 엑셀에서 한 번 살펴보자. 필터를 사용해 GDP per capita, PPP (current international $) 만 고른 뒤, 새로운 탭에 복사하여 붙여넣자. CSV 파일 형식으로 저장하자.

이 데이터를 Stata에서 불러오자(**[Stata 코드]** 참고).

- GDP는 문자열(string)이므로 이를 숫자형(numerical)으로 변환하자.
- 이 변수의 histogram을 그려보자. 이 그림은 어떤 시사점을 제공하나?
- CSV 파일을 Stata 파일로 새롭게 저장하자.

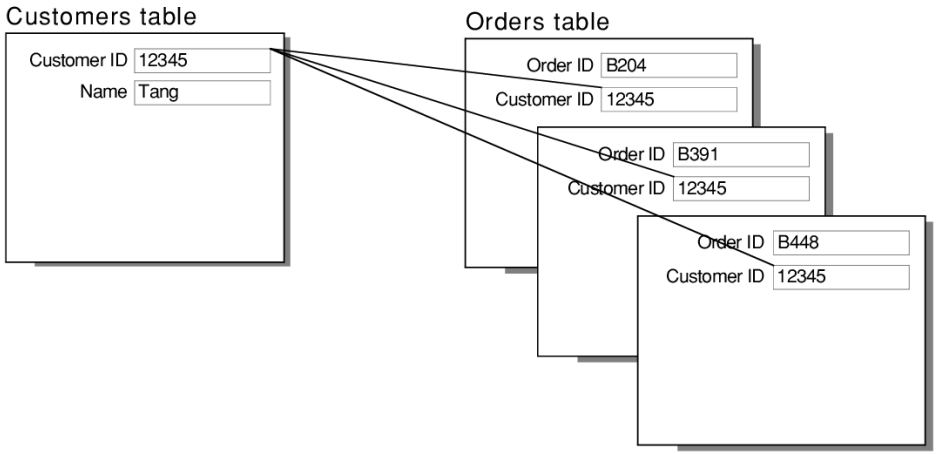
이 데이터는 있는 그대로(as it is)보다는 다른 서베이 데이터와 결합할 때 더욱 유용하다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

공공 데이터(Public Data)

두 데이터의 연결(merge)은 매우 중요한 자료 관리 테크닉 가운데 하나다 (**[Stata 코드]** 참고).

- Merge는 크게 네 가지 형식으로 가능하다: 일대일(one-to-one), 일대다(one-to-many), 다대일(many-to-one), 다대다(many-to-many)
- 각각 공통 변수(common identifier)가 "고유한 하나"인가 아니면 "중복된 여러 개"인가에 따라 다르다. 연구 목적에 따라서도 써야 하는 merge 방식이 다르다.
- Stata에서는 master 파일과 using 파일 모두에서 공통 변수의 이름이 반드시 같아야 한다.
- 공통 변수가 있는 한 수많은 데이터들을 연결할 수 있다. 연결된 데이터는 그렇지 않았을 때와는 비교될 수 없을만큼 큰 잠재력을 갖는다.



One-to-Many Merging

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

공공 데이터(Public Data)

이제 WVS 데이터와 WB 데이터로 돌아가서 두 데이터를 결합해 보자 (**[Stata 코드]** 참고).

- ISO 3166-1 국가 코드는 WVS 데이터와 세계은행 데이터 양쪽 모두에 들어있다. 이 공통 변수를 사용해 데이터를 연결(merge)할 수 있다.
- 먼저 WVS의 데이터에서 가치체계를 세 개의 더미변수(dummy variable)로 바꾸어보자.
- WVS의 데이터의 관찰단위(unit of observations)는 개인이였는데 이를 국가 수준으로 집계(aggregate)하자.
- 이제 WVS 데이터와 WB 데이터 모두 분석단위(unit of analysis)가 국가가 되었다. 서로 결합(merge)하자.
- 편의상 결합에 실패한 국가들은 그냥 삭제하자.
- 국가별로 물질주의적 가치 평균, 혼합형 가치 평균, 탈물질주의적 가치 평균이 GDP와 어떤 관련을 갖는지 산포도(scatterplot)를 그려보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

패널 데이터(Panel Data)

패널 데이터(Panel Data)

마지막으로 패널 데이터에 대해 살펴보기로 한다.

- 전통적인 사회통계학에서는 패널 데이터를 가장 우수한 데이터 형태로 여겨져 왔다.
- 패널 데이터의 수집과 관리가 대단히 까다롭다는 점을 감안하면, 대부분 대규모 조직이나 정부기관 등에서 특수한 목적에 따라 운용해 왔다는 사실은 놀랍지 않다.
- 대부분 패널 데이터의 질 관리는 매우 엄격하게 이루어진다. 하지만 패널은 응답자의 조사 거부, 이사 등 연락 두절, 사망 등의 이유로 어쩔 수 없이 마모(attrition)가 발생한다. 마모는 패널 데이터의 질에 부정적인 영향을 준다.
- 패널 데이터가 주어진 경우 좀 더 엄격하게 인과관계(causal relationship)에 관한 연구를 수행할 수 있는 여지가 생겨난다.
- 몇몇 연구자들(특히 석사학위 논문을 쓰려는 학생들)은 특정 연도의 자료 만을 잘라내 횡단면(cross-sectional) 자료로 삼아 분석하기도 했다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

패널 데이터(Panel Data)

몇몇 패널 데이터는 국내 사회과학자들에 의해 폭넓게 사용되어 왔다.

- 한국노동패널(노동연구원), 한국의료패널(한국보건사회연구원), 한국복지패널(한국보건사회연구원), 여성가족패널조사(한국여성정책연구원), 한국미디어패널조사(정보통신정책연구원), 청년패널조사(한국고용정보원), 한국교육종단연구(한국교육개발원), 가계금융복지조사(통계청), 장애인고용패널조사(한국장애인고용공단) 등 목적에 따라 다양하다.
- 많은 패널 데이터 학술대회에서는 거의 매해 컨퍼런스가 열릴때마다 "패널 데이터 분석방법론" 세션을 마련하고 있다. 아무 웹사이트나 가서 과거 컨퍼런스 일정표를 훑어보자.
- 모든 웹사이트에 가서 적어도 설문지(questionnaire)를 다운로드 받아 한 번씩은 살펴보아야 한다. 어떤 문항들이 있는지 눈여겨 보다보면 사회학적 상상력과 분석적 통찰력을 얻는다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

패널 데이터(Panel Data)

여기서는 한국교육종단연구 2005년 버전 자료를 살펴보기로 하자(**[Stata 코드]** 참고).

- 설문지(questionnaire)와 원시자료(raw data)를 본래 웹사이트에서 다운로드 받는 것이 제일이지만 패널 데이터의 경우 회원 가입부터 승인까지 꽤 시간이 걸리므로 교육용 데이터(KELS2005.zip)를 eCampus에서 다운로드 받자.
- 코드북을 보면 자기조절학습(self-regulated learning)의 측정문항 4개와 학업성취도(1학년 국영수) 총점 3개가 나온다. 이것들의 기술통계량(descriptive statistics)을 살펴보자.
- 자기조절학습의 문항들을 모두 더 해 하나의 복합변수(composite variable)를 만들고, 또 학업성취도 세 과목 총점의 복합변수를 만들자.
- 위 두 복합변수들의 히스토그램을 살펴보자.
- 두 변수 사이의 산포도(scatterplot)를 그려보고 약간 더 예쁘게 꾸며보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

Takeaways

Takeaways

데이터를 탐색하는데 시간을 충분히 많이 써야 한다.

- 찾고자 하는 데이터가 정말 존재하는지 그리고 어떻게 구하는지를 아는 것은 그 자체로 어려운 일이다(know-how라기보다 know-where에 가깝다).
- 자신의 전공과 무관하더라도 수많은 데이터를 살펴보아야 한다. 이 과정에서 새로운 분야에 대해 관심과 식견을 키우고 학제간 통찰력도 얻는다.
- 이론과 증거 사이의 간극은 이론적으로 매우 넓고 깊지만 현실적으로는 아주 사소하다. 얼핏 본 서베이 질문 하나로부터 지적 충격을 받아 완전히 새로운 이론적 관심으로 발전하는 일은 매우 흔하다.
- 책 읽기와 마찬가지로 코드북/설문지 읽기도 경험적 사회과학 연구에 큰 도움이 된다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY