계량분석
Examining Datasets

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

September 19, 2024

진행 순서

사회조사자료

사회조사자료

데이터 아카이브 안에서 수많은 사회조사자료를 찾을 수 있다.

- 사회조사(social survey)는 해당 지역 또는 국가의 거주민들이 정치·사회·경제·과학기술 등 여러 분야에 걸친 가치(values), 여론(opinions), 태도(attitudes), 행태(behaviors)를 조사하여 얻는다.
- 사회과학 분야에서 연구될 수 있는 여러가지 토픽을 커버한다. 시민 자유(civil liberties), 범죄/폭력, 그룹간 관용(intergroup tolerance), 도덕/윤리 판단, 국가재정, 심리적 안녕(psychological well-being), 사회적 계층 이동(social mobility), 스트레스와 트라우마 등 다양하다.
- 특정 토픽을 수 년에 한 번씩 돌리기도 한다.
- 물론 나이, 성별, 인종 등 기본적인 인구학적 변수(demographic variables)는 언제나 포함한다.

사회조사자료

- 일반적으로 하나의 사회조사자료는 특정 연구모집단 내 모든 구성원들에 대해 대표성 (representativeness을 확보하고 있다.
- (미국은 GSS, 한국은 KGSS, 중국은 CGSS, 일본은 JGSS, 독일은 GGSS 등) 국가별로 각각 대표적인 사회조사가 있다.
- 유사한 질문을 수 년에 걸쳐 반복적으로 질문하여 사회변동을 파악하기에 유용하다.
- 다만 같은 사람을 추적하지 않으므로 패널자료는 아니다. 다만 미국의 GSS는 패널도 함께 운용하고 있다.
- 데이터 수집 및 관리를 전공한 수많은 사회학자들이 프로젝트를 운영하기 때문에 일반적으로 대단히 품질이 높다.

사회조사자료

사회조사 데이터들을 비교하여 국가간 비교를 수행할 수도 있다.

- 예전 비교사회학(comparative sociology)이 제법 흥성하던 시기가 있었다.
- 이 목적으로 설계·수집된 대표적인 자료로 Ronald Inglehart가 주도한 World Values Survey (WVS)가 특히 유명하다.
- 유럽 국가들 사이에서 설계·수집된 자료로 European Values Survey (EVS)도 나름 유명하다.

사회조사자료

- 몇몇 국가의 연구자들은 자국에서 나름의 GSS를 운영하면서 같은 모듈(module)을 같은 해에 함께 질문하여 일부러 비교가능하도록 설계하였다.
- 이 자료를 따로 뽑아 International Social Survey Programme (ISSP)을 구축했다.
- 많은 유럽 국가들은 ESS (European Social Survey)에 함께 참여한다.
- 동아시아의 국가들 간에 운영되는 East-Asian Social Survey (EASS)도 있다.
- (필요에 따라) 한국인과 미국인의 사회적 가치를 비교하기 위해 KGSS와 GSS를 함께 분석할 수 있다.

사회조사자료

- 오늘날 비교사회학은 크게 침체되었지만, 폭넓은 자료 이용가능성(data availability)에 힘입어 다양한 주제를 탐구하며 나름의 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

사회조사자료

세계가치관조사의 최근 자료를 간단히 살펴보자.

- 링크는 https://www.worldvaluessurvey.org
- Stata 원자료 파일(Wave 7) 뿐만 아니라 설문지도 함께 다운로드 받자.
- 특히 탈물질주의적 가치(post-materialist values)를 살펴보자.
- 다음의 중 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
- Y002 변수가 어떻게 만들어졌는지 확인해보자.

잉글하트, 로널드. 1983.『조용한 혁명』. 종로서적.

잉글하트, 로널드 · 크리스찬 웰젤. 2011.『민주주의는 어떻게 오는가』. 김영사.

패널자료

패널자료

패널자료의 분석은 많은 잠재력을 내포한다.

- 하지만 현실적으로 패널에는 응답자의 조사 거부, 이사 등 연락 두절, 사망 등의 이유로 어쩔 수 없이 마모(attrition)가 발생한다(마모는 당연히 자료의 질에 부정적인 영향을 준다).
- 패널자료의 질 관리는 매우 어렵고 또한 엄격하게 이루어져야 한다.
- 패널자료의 수집과 관리가 대단히 까다롭다는 점을 감안하면, 대부분 대규모 조직이나 정부기관 등에서 특수한 목적에 따라 운용해 왔다는 사실은 놀랍지 않다.

패널자료

- 전통적인 사회과학 통계학에서는 패널자료를 가장 이상적인 자료형태로 여겨왔다.
- 패널자료 분석은 훨씬 고난이도의 분석기법을 요구한다.
- 가령 패널자료 분석을 통해 관찰자료인 경우에도 엄격한 인과관계(causal relationship) 연구를 수행할 수도 있다.
- 몇몇 연구자들(특히 석사학위 논문을 쓰려는 학생들)은 패널자료 전체를 분석하기 보다 특정 연도의 자료만을 잘라내 횡단면 자료로 삼아 분석하기도 한다.

패널자료

국내만 해도 수많은 패널 데이터가 존재한다.

- 한국아동청소년 패널조사, 한국청소년 패널조사, 다문화청소년 패널조사, 학업중단청소년 패널조사(이상 한국청소년정책연구원)
- 서울교육종단연구, 서울교원종단연구(이상 서울교육정책연구소)
- 사업체패널, 노동패널(이상 노동연구원)
- 한국의료패널, 한국복지패널(이상 한국보건사회연구원)
- 인적자본기업패널, 한국교육고용패널(이상 한국직업능력개발원)
- 여성가족패널조사(한국여성정책연구원), 한국미디어패널조사(정보통신정책연구원), 청년패널조사(한국고용정보원), 한국교육종단연구(한국교육개발원), 가계금융복지조사(통계청), 장애인고용패널조사(한국장애인고용공단), 한국아동패널(육아정책연구소), 청소년건강행태조사(질병관리청) 등

공공 데이터

공공 데이터

사회학 연구에서 정부의 공식 통계는 다소 미묘한 입장에 있다.

- Emile Durkheim의 〈자살론〉은 공식 통계(official statistics)를 활용한 가장 뛰어난 사회학 고전 연구다(동시에 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?"
- 아까 언급한 행정자료(administrative data)도 어떤 의미에서 여기와 잇닿아 있다.

공공 데이터

- 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 정보공개 청구로 확보할 수 있는 경우가 있다.
- 공공 데이터의 원자료가 바로 행정자료인 경우가 많다(Why?).

공공 데이터

근래에는 공공 데이터의 양과 범위가 점점 넓어지고 있다.

- 이른바 4차 산업혁명의 한 인프라로서 공공 데이터의 가치가 재발견되면서 연구 기회도 늘어났다.
- 각종 경진대회가 열리기도 한다(예컨대 2023년 문화공공데이터 활용 경진대회).
- 비정형인 데이터가 일반화되면서 좀 더 새롭고 창의적인 접근이 요구되는 경우가 많아졌다.
- 다운로드가 아니라 Application Programming Interface (API)의 형식을 취하는 경우가 늘어났다.
- 이런 경우에는 자료에 접근하기 위해서라도 코딩(주로 R이나 Python)을 배워야 하고 전처리 숙련도도 필요하다.

공공 데이터

여기서는 World Bank에서 제공하는 공공 데이터를 살펴보자.

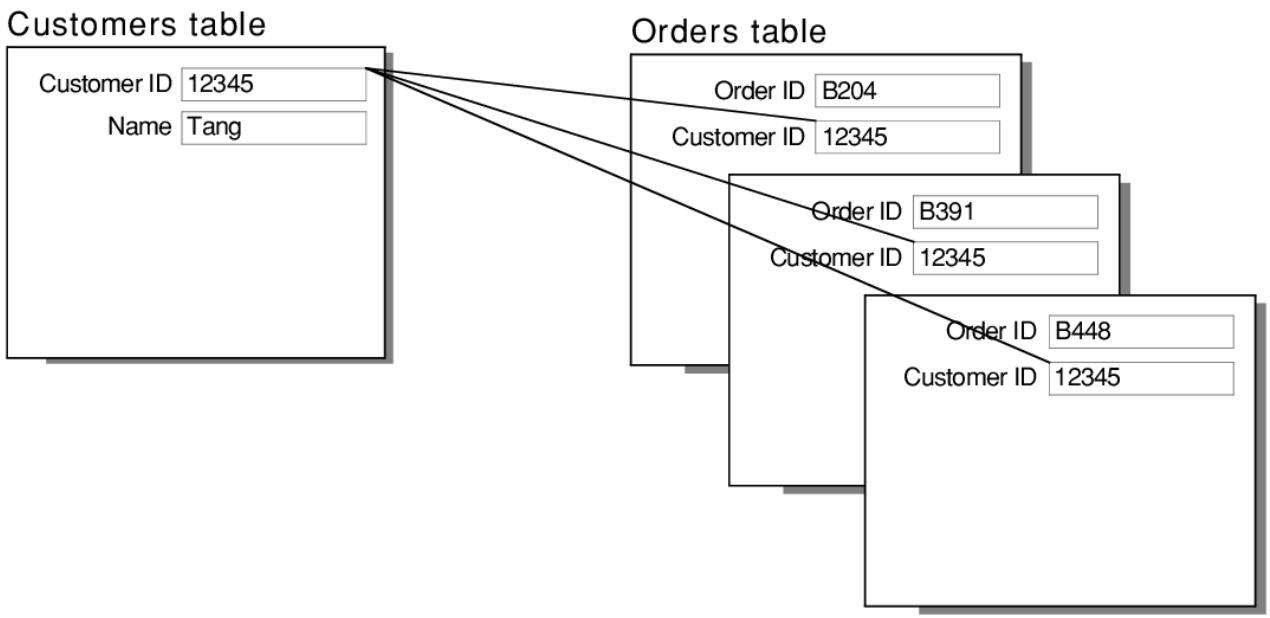
- 세계은행(https://data.worldbank.org)에서 국가 단위의 경제 통계를 다운로드 받을 수 있다.
- World Development Indicators 메뉴를 선택하면, 모든 Country, 모든 Series, 2019년을 체크해서 국가 단위의 경제 통계를 csv 파일의 형식으로 다운로드 받을 수 있다.
- GDP는 문자열(string)이므로 이를 수치형 자료로 변환하고 히스토그램을 그려보자. 이 그림은 어떤 시사점을 제공하나?

자료의 결합

자료의 결합

두 개 이상의 자료의 연결은 매우 중요한 기법이다.

- 자료 연결(data merging)는 크게 네 가지 형식 중 하나로 이루어진다:
  (1) 일대일(one-to-one)
  (2) 일대다(one-to-many) (3) 다대일(many-to-one)
  (4) 다대다(many-to-many)



One-to-Many Merging

자료의 결합

- 각각 공통 변수(common identifier)가 “고유한 하나”인가 아니면 “중복된 여러 개”인가에 따라 다르다.
- 연구 목적에 따라서도 써야 하는 연결 방식이 달라진다.
- Stata에서는 결합하는 파일(master file)과 결합에 쓰일 파일(using file) 모두에서 공통 변수의 이름이 반드시 같아야 한다.
- 공통 변수가 있는 한 수많은 자료들을 연결할 수 있다. 연결된 자료는 그렇지 않았을 때와는 비교할 수 없을만큼 큰 잠재력을 갖는다(Why?).
- 특히 공공 데이터는 그 자체로는 다소 쓸모가 약하지만, 다른 사회조사 자료나 패널자료와 결합되었을 때 나름의 가치를 보여준다(Why?).

자료의 결합

아까 준비한 세계가치관조사 자료와 세계은행 공공 데이터를 결합해 보자.

- ISO 3166-1 국가 코드는 WVS 자료와 세계은행 공공 데이터 양쪽 모두에 들어있다. 이 공통 변수를 사용해 자료를 연결할 수 있다.
- 먼저 WVS 데이터에서 가치체계를 세 개의 더미변수(dummy variable)로 바꾸어보자.
- WVS 원자료의 관찰단위는 개인이였는데 이를 국가 수준으로 집계(aggregate)하자.
- 이제 WVS 데이터와 WB 데이터 모두 측정단위는 국가가 되었다.
- 이제 서로 결합(merge)하고 편의상 결합에 실패한 국가들은 그냥 삭제하자.
- 국가별로 물질주의적 가치 평균, 혼합형 가치 평균, 탈물질주의적 가치 평균이 GDP와 어떤 관련을 갖는지 산점도를 그려보자.

시사점

시사점

책 읽기와 마찬가지로 설문지 읽기도 사회과학 연구에 큰 도움이 된다.

- 찾고자 하는 데이터가 정말 존재하는지 그리고 어떻게 구하는지를 아는 것은 그 자체로 어려운 일이다.
- 이런 것들은 노하우(know-how)라기보다 노웨어(know-where)에 가깝다.
- 자신의 전공과 무관하더라도 수많은 데이터를 살펴보아야 한다.
- 이 과정에서 새로운 분야에 대해 관심과 식견을 키우고 학제간 통찰력도 얻는다.
- 얼핏 본 사회조사 설문지의 질문 하나로부터 지적 충격을 받아 완전히 새로운 이론적 관심으로 발전하는 일은 매우 흔하다.
- 최대한 많은 설문지들을 다운로드 받고 어떤 문항들이 있는지 눈여겨 보다보면 사회학적 상상력과 분석적 통찰력을 얻는다.

시사점

자료 탐색에 시간을 충분히 많이 써야 한다.

- 시간을 많이 들여서 관심에 부합하는 데이터 아카이브를 발굴하고 자주 살펴보는 습관이 필요하다.
- 평상시에 수많은 자료에서 조사하는 변수들을 미리 머리 속에 잘 정리해두자.
- 이론 또는 가설이 생겨났을 때 “아! 이 가설이라면 세계가치관조사(WVS)가 적절하겠구나!” 하고 깨닫는 것이 보다 이상적이다.
- 많은 패널자료 학술대회에서는 거의 매해 컨퍼런스가 열릴때마다 “패널 데이터 분석방법론” 세션을 마련하고 있다(아무 웹사이트나 가서 과거 컨퍼런스 일정표를 훑어보자).

연습문제

연습문제

연습 1. 한국교육종단연구 2005년 교육용 자료(Y2_STD_EDU.SAV)를 사용하여 다음 지시를 수행하시오.

- 이 SPSS 자료를 Stata로 불러오시오.
- 자기조절학습(self-regulated learning)의 측정문항 4개와 학업성취도(1학년 국영수) 총점 3개의 기술통계(descriptive statistics)를 살펴보시오.
- 자기조절학습의 문항들을 모두 더하여 하나의 복합변수(composite variable)를 만드시오.
- 마찬가지로 학업성취도 세 과목 총점의 복합변수를 만드시오.
- 위 두 복합변수들의 히스토그램을 살펴보시오.
- 두 변수 사이의 산점도를 그려보고 약간 더 예쁘게 꾸미시오.

연습문제

연습 2. 같은 자료로 다음 지시를 수행하시오.

- 사회 자기 개념 변수들을 모두 더하여 사회 자기 총점을 계산하시오.
- 사회 자기 점수를 각 학교별(schid)로 집계화한 평균값을 구하여 그 결과를 적절히 시각화하시오.
- 각 학교별로 집계된 사회 자기 점수의 평균값을 원래 자료(Y2_STD_EDU.SAV)에 적절히 결합하시오. 결합의 과정을 설명하시오.
- 결합된 자료에서 학생 개인의 사회 자기 점수와 학교별 평균값의 편차(deviation)를 계산하고, 그 결과를 시각화하시오.