계량분석
강의 소개

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

September 5, 2024

진행 순서

1. 강사 소개
2. 사회학에서 계량분석
3. 수업의 구성과 당부사항
4. 통계분석 패키지의 선택
5. 충북대 사회학과 커리큘럼과 이 수업

강사 소개

강사 소개

김현우

- Pennsylvania State University 사회학 박사학위 취득(2017)
- Edna Bennett Pierce Prevention Research Center 박사후 연구원 (2017-2019)
- Edna Bennett Pierce Prevention Research Center 연구조교수 (2019-2021)
- 충북대학교 사회학과 조교수 (2021.9-)

사회학에서 계량분석

사회학에서 계량분석

오늘날 통계적 리터러시가 갖는 중요성이 폭발적으로 증대하였다.

- 산업계와 학계에서 제4차 산업혁명의 기반기술(빅데이터, 인공지능, IoT 등)에 대한 전문가 수요는 폭발적이다.
- 단순히 저숙련자의 수요는 이제 사그러드는 분위기로 고숙련자의 수요가 높아지고 있다.
- 생산되는 데이터의 양도 급속히 증가하고 있다.
- 통계학은 인공지능(AI), 빅데이터 분석, 기계학습(machine learning) 등의 기초가 된다.

사회학에서 계량분석

사회학에서 (고급) 계량적 분석은 이미 핵심적 주류다.

- ASR 2024년 8월호에 실린 6편의 논문 중 5편의 논문이, AJS 2021년 9월호에 실린 4편의 논문 중 2편의 논문이 계량적 분석을 사용하였다.
- 여러분의 세부전공과 상관없이 만약 계량분석을 공부하지 않으면 $70\%(=7/10)$는 읽고 비판하지 못한다.
- 최근 10년 사이 사회학 방법론 분야에도 눈부신 발전이 있었다.
- (만약) 계량적 분석을 택하여 공부한다면 이를 필수적으로 따라가야 한다.
- 여러분의 졸업 이후에도 이 발전은 (더 빠르게) 계속 될 것이며 스스로 그 지식을 업데이트할 수 있어야 한다.

사회학에서 계량분석

"에이, 그래도 나는 계량분석을 전공할 건 아닌데..."

- 물론 모든 사람이 계량적 분석 자체를 전공해야 하는 것은 아니다.
- 하지만 이 수업에서 다루는 토픽은 선형회귀분석(linear regression analysis)의 기초와 그 응용으로, 이는 사회학에서 쓰이는 모든 계량분석의 기초적인 뼈대를 구성한다.
- 다시 말해, 이 수업이 다루는 토픽은 기초일 뿐이고 전혀 수준높은 계량분석이 아니다.

사회학에서 계량분석

회귀분석에도 두 개의 얼굴이 있다.

- 과학(science)으로서의 측면과 손기술(art)로서의 측면이 있다.
- 과학은 결국 설명할 수 있는 힘이다.
- (1) 내가 지금 하고 있는 것이 무엇인지 직관적으로 명확하게 알아야 한다.
- (2) 남에게 그리고 논문 속에서 자신의 분석을 확실하게 설명할 수 있어야 한다.
- 손기술은 반복숙련의 힘이다.
- (1) 손기술/연습은 생각보다 매우 중요하며 연습해두지 않으면 연구나 실무에서 전혀 쓸 수 없다.
- (2) 투자한 시간이 절대적으로 많아야 한다. 머리 못지 않게 엉덩이로 공부한다!

사회학에서 계량분석

수학을 제대로 하지 않고 건너뛴다는 것은 사실 넌센스다.

- 통계학은 본디 수학의 분과학문이다.
- 강의계획서와 강의안을 만드는 동안에도 사실 몇 번이나 마음을 바꾸었다. 여러 다른 교수님들과도 상의를 해보았다.
- 최종적으로 수학을 필요최소한만 하고 모두 건너뛰기로 하였다. 수학보다 직관을 강조하기로 한 셈이다.
- 수학을 건너뛰는 것에는 장단점이 있다. 중요한 장점은 사회통계학에 대한 흥미를 잃지 않게 돕는다는 것이다.
- 그러므로 지금은 일단 직관적으로 공부한 다음, 나중에 점점 더 호기심이 생겨 계량분석을 한층 깊게 공부하고 싶다면 그때 가서 수학을 하자!

수업의 구성과 당부사항

수업의 구성과 당부사항

교재는 지갑 사정이 허용하는 한 많이 구입한다.

- 기초통계 교과서는 여러 권이 필요할 수도 있다. 주교재는 반드시 구입하여 연습문제를 과제로 풀어야 한다.
- 회귀분석의 원리가 강조된 책은 구입하는 것이 좋다.
- Stata 책은 빌려도 된다. 이런 것들은 소모품에 가깝다.
- 정확히 알고 구입하는 것이 아니라면 수리통계학과 같이 지나치게 수학적인 것은 피할 것.
- 전공은 전혀 상관없다(e.g., 경영통계학, 관광통계학 등).

수업의 구성과 당부사항

모르는 것이 있을때는?

- 구글(Google), 유튜브(YouTube), 스테이타리스트(Statalist), 스택오버플로(Stack Overflow), 챗GPT 등 인터넷 자료는 강력한 우군이다.
- 인터넷을 찾아보아도 영 답이 안나온다 싶으면 바로 달려와서 연구실 문을 두드릴 것.
- 피해야 할 것들이 있다:

  (1) 쪽팔림/죄책감으로 인해 교수자에게서 도망쳐 다니는 것.

  (2) 장문의 메일로 묻는 것. 무슨 소린지 이해를 못할 때가 많고, 설령 알아들어도 메일로 답하기 곤란할 때가 많다.

수업의 구성과 당부사항

교실 밖에서도 여러분이 해야 할 일이 있다.

- 퀴즈는 정말로 여러분을 위한 것! 나도 몹시 귀찮다.
- 중간시험과 기말시험에서 수업 시간에 학습한 기법이 담긴 기존 문헌 3편을 비판적으로 리뷰한다(중간시험 3편, 기말시험 3편).
- 반드시 등재지 혹은 SSCI 수준의 학술 논문을 선택하여, 여러분의 논문에 활용될 수 있는 형태로 요약한다.

수업의 구성과 당부사항

수업에서 분석할 데이터를 준비하자.

- 이 수업을 가이드로 삼아 여러분은 스스로 관심있는 프로젝트를 직접 수행해보는 것이 좋다.
- 첫 출발점은 분석할 데이터를 준비하는 것!
- 먼저 관심가는 주제(경제/환경/조직/노동/젠더 등)를 먼저 특정할 것. 그리고 나서 데이터를 고른다. 반드시 한국 데이터가 아니어도 된다.

수업의 구성과 당부사항

어디에서 데이터를 확보할까?

- ICPSR: https://www.icpsr.umich.edu
- KGSS: http://kgss.skku.edu
- GSS: https://gss.norc.org
- KSDC: https://www.ksdc.re.kr
- 사업체패널: https://www.kli.re.kr/wps
- 한국노동패널: https://www.kli.re.kr/klips

수업의 구성과 당부사항

데이터 확보는 제법 중요한 일!

- 좋은 데이터 찾기에 시간을 많이 쏟아야 한다. 좋은 데이터를 얻는 것은 좋은 논문의 출발점.
- TITO (Trash In, Trash Out).
- 지도교수님과 반드시 이런 이야기를 나누고 조언을 받을 것!
- 이상적으로는 이 수업에 좋은 데이터를 들고와서 석박사 논문을 위한 분석을 바로 하는 것!
- 만일 논문 주제를 정하지 못했거나 데이터를 찾지 못할 것 같다면 석박사 논문과는 다른 주제의 (관심가는) 데이터라도 상관없다.

수업의 구성과 당부사항

영어는 적극 익히고 사용해야 한다.

- 처음엔 좀 어색할 수도 있지만, 통계학은 결국 영어권에서 그 핵심적인 발전이 이루어진 분야이므로 좋건싫건 이것이 오리지널이다.
- 자꾸 우리말과 엮어서 반복할 예정이지만, 결국엔 둘 다 기억해야 한다.
- 게다가 우리말이라고는 해도 순우리말이 아니라 일본식 한자가 많다.
- 영어가 훨씬 직관적이라 통계학 속에 담긴 진짜 의미나 취지를 오히려 알아듣기 쉽다.
- 필요하다면 한국통계학회 통계용어집을 이용할 것.

통계분석 패키지의 선택

통계분석 패키지의 선택

사회과학자들 사이에서는 크게 다섯 개의 대안들이 알려져 있다.

1. SPSS (Statistical Package for Social Sciences)
2. SAS (Statistical Analysis System)
3. Stata
4. R
5. Python

통계분석 패키지의 선택

가격 측면

- SPSS와 SAS는 일단 아웃. 개인은 도저히 감당할 수 없고 반드시 기관이 구입해야 한다.
- Stata는 개인이 살짝 부담이긴 한데 살 수는 있다. 물론 기관이 부담없이 구입해 주기에도 좋다.
- R과 Python은 무료이므로 이 측면에서 압도적이다.

통계분석 패키지의 선택

학습 난이도 측면

- R과 Python은 학습 곡선(learning curve)가 가파르다. 언어 자체를 배우는데 시간을 꽤 써야한다. 일단 습득하면 가장 많은 것들(머신러닝, 웹스크래이핑, 빅데이터, GIS, SNA, 베이즈 통계 포함)을 할 수 있다!
- Stata와 SAS는 학습 곡선이 살짝 있지만 한 학기 안으로 할 만하다. 요즘엔 베이즈통계나 머신러닝 등을 조금 다룰 수 있지만, R과 Python에 비한다면 한참 멀었다.
- SPSS는 학습 곡선이 낮다. 구문(syntax)을 배우려면 약간 시간이 걸리긴 하지만 거의 쓸데가 없다. 고급통계분석에는 제약이 많다.

통계분석 패키지의 선택

인기 측면

- SAS는 미 정부에서 여전히 압도적이다. 최근에 좀 바뀐다는 기류가 있지만 아직 잘 모르겠다.
- SPSS는 미국 대학 학부와 국내 조사기관에서 압도적인 인기를 누리고 있다. 저숙련자 사이에서는 거의 지배적이다.
- R과 Python은 전세계적으로 개발자 커뮤니티에서 압도적이다. 단순 규모로서는 최대 인기!
- Stata는 미국 의료기관(생물통계)과 국내외 사회과학 대학원에서 상당한 인기를 누리고 있다.

통계분석 패키지의 선택

왜 결론은 Stata인가?

- 일단 SAS는 내가 모른다.
- SPSS는 현업(조사기관)에서 많이 쓰이지만 그 밖에서는 잘 안쓰인다. 게다가 통계학을 알면 쉽게 독학할 수 있다(jamovi 포함).
- Python과 R은 (머신러닝, 웹스크래이핑, 빅데이터 등에서 강력하므로) 앞으로 취업을 앞둔 학생들을 위해서 이쪽의 전망이 낫지만 두 개의 단점이 있다.

  (1) 학습 곡선 때문에 포기자가 나온다(회귀분석도 배우기 어려운데 이것까지 익히느라 부담이 높아진다).

  (2) 분석에 소모되는 반복 작업에서 Stata가 훨씬 더 편리하다.
- 미국 Penn State, UPenn, Notre Dame 등에서도 Stata가 널리 채택되어 쓰인다 (보통 1학점 짜리 랩에서 공부하는 경우가 많다).

충북대 사회학과 커리큘럼과 이 수업

충북대 사회학과 커리큘럼과 이 수업

외부 특강과 비교해보자.

- 많은 학생들이 통계분석 패키지를 공부하고 논문 쓰기 위해서 대학 바깥에서 특강을 듣는다.
- KOSSDA 외부 특강(https://kossda.methods.snu.ac.kr/) 등이 대표적이다.
- 이 수업은 특히 “기초통계학” 및 “중급통계학”과 같은 수요에 대응한다.
- 외부 특강의 수강료는 약 60만 원 정도 한다.

충북대 사회학과 커리큘럼과 이 수업

충북대 사회학과 학부에서 개설하는 계량분석 관련 수업

- 사회통계(2학년): 통계학의 논리적 기초 중심. Excel.
- 사회통계연습(2학년): 실습 중심. SPSS.
- 사회조사방법론(3학년): 소논문 작성. SPSS.
- 소셜데이터사이언스(4학년): 파이썬 입문, 기계학습, API (Applicable Programming Interfaces), 소셜네트워크 분석(social network analysis)

충북대 사회학과 커리큘럼과 이 수업

충북대 사회학과 대학원에서 개설할 수 있는 관련 수업

- 계량분석: 이 수업
- 범주형 자료분석(categorical data analysis). 대학원 계량분석이 선수강.
- 고급사회통계세미나: 수요에 따라 종단자료 분석(longitudinal data analysis), 다층모형 분석(multilevel modeling), 사건사분석(event-history analysis), 비실험적 인과분석(causal inference with observational data). 범주형 자료분석이 선수강.
- 소셜데이터사이언스: 소셜네트워크분석(social network analysis), 지리정보시스템(Geographical Information System), 텍스트 분석, 기계학습.

충북대 사회학과 커리큘럼과 이 수업

충북대 사회학과에서 개설되지 “않는” 계량분석 관련 수업들

- 데이터베이스: 데이터의 보관 및 접근. DBMS (database management system 설계 및 구축(SQL 포함), 클라우드 컴퓨팅(cloud computing) 등 데이터 엔지니어링 이슈 등.
- 서베이 방법론: 확률표집원리, 조사 및 문항설계, 조사품질관리, 패널관리, 가중치 및 결측치 핸들링, 온라인 서베이 등