SFERS DATA Seminar with Python

Yitae Kwon

SFERS of SNU

2024-1

Goals

1. 데이터 분석 도구들과 친해지기
2. 데이터를 직접 수집하고 가공하기
3. 데이터를 요약하고 시각적으로 표현하기
4. 계량경제학 방법론을 이론적으로 이해하기
5. 통계적 결정이론/검정이론을 이해하기
6. 데이터 특성에 맞는 분석 방법과 모형을 결정하기
7. 실제 데이터를 이용하여 계량분석을 수행하기
8. 모형의 적합성을 판단하기
9. 분석 결과를 해석하고 정리하기
10. (Optional) 시뮬레이션과 수치실험 해보기

What we will and won't

- **Will**
  - 간단한 수리, 통계적 지식에 기반한 모형 소개
  - Python 기초 문법 공부, 이미 만들어진 함수들 가져다 쓰기
  - 전통적 통계학 기반의 '추정' 및 '검정' 방법론들
  - Jupyter Notebook, Overleaf 그리고 Github 사용하기
  - 간단한 모형과 작은 데이터 이용하기
  - 대충대충 넘어가기, 수강생에게 떠넘기기
  - O/X 위주의 질의응답
- **Won't**
  - 과도한 수리적, 통계적 계산
  - 과도한 프로그래밍 공부, R 문법 공부
  - 머신러닝/딥러닝 기반의 '예측' 방법론들
  - Word, PowerPoint, 그리고 카카오톡 파일공유 사용하기
  - 복잡한 모형과 큰 데이터 이용하기
  - 하나하나 파고들기, 책임지기
  - 대신 분석하기

Our Languages

- Natural Language: 한국어
- Programming Language: **Python3**(세부버전 무관; 저는 3.11.5긴 함)
- File Type: `.csv`(data), `.ipynb`(code), `.tex`(paper/slides)
- IDE: Excel(code), Visual Studio Code(code), Overleaf(paper/slides)
- OS: Windows10
- Platform: GitHub(Web/GitHub Desktop)
- Mathematical Language: 경제수학, 경제통계학은 적어도 수강
- Economical Language(?): 무관
- Programming Experience: 적어도 시도해본 경험은 있어야 함
  $\Rightarrow$ 없어도 되기는 하나 단기간에 배우기 고통스러울 것

Schedule

전반부 일정

- **Tools for Data Science**
  - **Installing** Python, VS Code and GitHub Desktop (HW[1])
- **Data Collection**
  - Lec1(03/19): **Crawling** (IP[2])
- **Data Wrangling**
  - Lec1/2(03/19&26): **Wrangling** (IP, HW)
- **Data Presentation**
  - Lec2(03/26): **Visualization** (IP, HW)
  - Lec2(03/26): **Summarizing** with LaTeX (IP, HW)
  - Ensuring data reproducibility with GitHub (HW)
- **Data Analysis(1): Regression**
  - Lec1/2/3(04/02): **Regression** (IP, HW)

[1] HomeWork

[2] In Person

Schedule

후반부 일정

- **Data Analysis(2): Traditional Econometrics**
  - Lec4(04/09): **Experiment** and **Causal Inference** (IP, HW)
  - Lec5(04/30): **Time Series Analysis** (IP, HW)
- **Data Analysis(3): Data Science**
  - Lec6(05/07): **Dimension Reduction** and **Nonparametric Stats** (IP)
  - Lec7(05/14): **Classification** and **Clustering** (IP)
  - Lec8(05/21; Optional): **Simulation** and **Optimization** (IP? HW?)

In-Person Lectures and Homeworks

- 시간: 화요일 18:00∼20:00
  ⇒ 저도 시간 긴 거 안 좋아하기는 한데 양이 많아서...
- 장소: 관정 스터디룸(대면)
  ⇒ 예약해주실 분 구함
- 세미나 방식: 강의 + 실습
  ⇒ 이론은 강의만 듣고, 실습은 직접 Python으로 해볼 예정
- 숙제: 준비물/강의후 실습
  ⇒ 프로그램 설치, 데이터 분석 과정 재현, 보고서 작성 등
  ⇒ 필수는 아니나 적극 권장(이왕 들으시는거...)
- 주의: 오타/오류 많을 수 있음(양해 부탁)

One-Page Summary of Topics

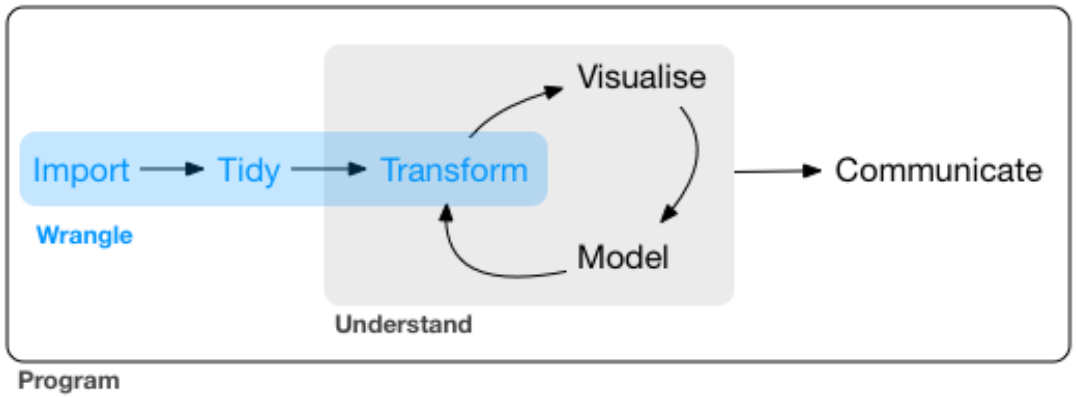
Lec1. Crawling

- **예상 사용 패키지**: `Selenium`, `Requests`, `Pandas`
- **경제학자(researcher)** vs. 데이터과학자(technician)
- 표본 설계, 실험 계획, 데이터 수집, 변수 설정은 해당 분야의 지식(domain knowledge)을 가진 경제학자가 주도적으로 수행해야 함
  $\Rightarrow$ 이 과정을 모두 분석가에게 맞기면 경제학자의 역할은?
- 다행히도 표본 설계, 실험 계획, 서베이/실험은 돈으로 해결 가능
- 적어도 **데이터 수집**은 직접 해보자! (학부생은 돈이없다)
- 클릭하여 사진 혹은 데이터를 모아보자 $\Rightarrow$ `Selenium`
- 파일을 자동으로 다운로드받자 $\Rightarrow$ `Requests`
- 파일을 웹/로컬에서 읽거나 쓰자 $\Rightarrow$ `Pandas`
- **도움되는 팀**: 금융팀(데이터수집)

One-Page Summary of Topics

Lec1/2. Wrangling

- **예상 사용 패키지**: `Pandas`, `Numpy`
- 파이썬이 데이터과학의 맹주가 되게 만들어준 1등공신들
- 일반적으로 우리가 얻은 데이터는 굉장히 더럽다(raw/untidy).
- 시각화, 모형화, 해석을 위해서는 깔끔한 데이터가 필요하다(tidy)[3].
- **도움되는 팀**: 금융/계량팀(데이터정리)



[3]아래의 그림은 R for Data Science에서 가져왔다.

One-Page Summary of Topics

Lec1/2/3. Regression

- **예상 사용 패키지**: `statsmodels`
- 변수 간의 관계를 밝히고, 통계적으로 유의한지 확인하는 도구
- 선형적이고, 가법적인 모형을 가정하기에 간단하고 해석하기 편함
- 상관관계(association)일 뿐, 인과관계(causation)는 아님
  $\Rightarrow$ 내생성(endogeneity) 문제 등이 존재할 수 있음
- **통계적 방법론**: 단순선형회귀, 다중선형회귀, 가변수(더미변수), 다항회귀, 스플라인회귀, 모형진단, 일반화선형모형, 일반화가법모형
- **참고 페이지**: An Introduction to Statistical Learning with Python
- **도움되는 팀**: 금융1팀(panel data), 계량1팀(regression tree)

One-Page Summary of Topics

Lec4. Experiment and Causal Inference

- **예상 사용 패키지**: `statsmodels`, `linearmodels`
- 변수 간의 인과관계를 밝혀 보자
- 과학적 실험에서의 인과관계 확인
- 자연실험과 준실험에서의 인과관계 확인
- **통계적 방법론**: t-검정, ANOVA, 순열검정, 콜모고로프-스미르노프 검정, IV/2SLS/IV-GMM, DiD, Synthtic Control, RDD, 매칭, 웨이팅
- **참고 페이지**: Causal Inference for the Brave and True
- **도움되는 팀**: 금융1팀(panel data), 계량1팀(causality)

One-Page Summary of Topics

Lec5. Time Series Analysis

- **예상 사용 패키지**: `statsmodels`
- 변수 간의 시계열적 인과관계를 밝혀 보자
- 두 요소의 시간적 관계를 확인하는 것도 인과성 분석의 일종
- 단변량 시계열 혹은 다변량 시계열에서의 시계열분석
- 확률론에 대한 논의 없이 진행할 예정
- **통계적 방법론**: SARIMA, ARIMAX, GARCH, Decomposition, VAR, VECM, Granger Causality, Change-Point Tests, Quantile Regression
- **참고 페이지**: Forecasting: Principles and Practice, Machine Learning Tutorial, Time Series Analysis in Python with statsmodels
- **도움되는 팀**: 금융1팀(panel data)

One-Page Summary of Topics

Lec6. Dimension Reduction and Nonparametric Models

- **예상 사용 패키지**: `statsmodels`, `scikit-learn`
- 데이터과학의 시대! 고차원 데이터에서의 분석
- 다양한 형태의 데이터를 분석하기 위한 비모수적 방법들
- 관계를 확인(confirmatory)하기보다는, 관계를 탐색(exploratory)
- **통계적 방법론**: Ridge, Lasso, PCA, FA, Bootstrap, KDE, Nonparametric Regression, Multiple Comparisons(FDR, FWER)
- **참고 페이지**: An Introduction to Statistical Learning with Python
- **도움되는 팀**: 금융1팀(multiple comparison), 계량2팀(bootstrap)

One-Page Summary of Topics

Lec7. Classification and Clustering

- **예상 사용 패키지**: `statsmodels`, `scikit-learn`
- 관계 분석보다는 어떠한 task를 수행하는 ML 방법론들 소개
- 가장 대표적인 task인 classification과 clustering 방법론만 간단히 소개
- DL은 제가 선호하지 않는 관계로 알아보지 않습니다.
- **통계적 방법론**: Linear Discrimination, KNN, Tree-Based Methods, SVM, K-Means Clustering, Hierarchical Clustering, DBSCAN
- **참고 페이지**: An Introduction to Statistical Learning with Python
- **도움되는 팀**: 계량1팀(tree-based methods)

One-Page Summary of Topics

Lec2. Visualization and Summarizing

- **예상 사용 패키지**: `matplotlib`, `seaborn`
- 우리가 앞서 wrangling한 데이터들, 모형을 통해 분석한 결과들을 어떻게 표현해야 할까?
- 표와 그래프를 통한 시각화 ⇒ `matplotlib`, `seaborn`
- 코드를 정리하기 ⇒ Jupyter Notebook
- 글과 슬라이드를 이용한 정리 ⇒ LaTeX
- 이러한 과정은 데이터 분석에서 필수적
  ⇒ 실제로는 가장 마지막 단계이지만, 가장 먼저 배울 것
- **도움되는 팀**: 모든 팀

One-Page Summary of Topics

Lec8. Simulation and Optimization

- **예상 사용 패키지**: `PyTorch`, `NumPy`
- 너무 소외되는 것 같은 미시팀과 금융2팀을 위한 주제
- 수치적인 시뮬레이션, 그리고 컴퓨터를 통한 최적화 문제의 풀이법
- 손으로, 수리적으로 풀 수 없는 문제를 컴퓨터로 풀자
- 수치선형대수, 수치해석개론, 고급통계계산 내용 일부
- **통계적? 방법론**: Monte Carlo Simulation/Integration, RNG, 수치미적분, Cholesky decomposition, Newton's Method(Root finding), Runge-Kutta 4th order method(ODE/PDE), Optimization
- **참고 페이지**: 원중호 교수님 강의노트, 신동우 교수님 강의노트
- **도움되는 팀**: 미시팀, 금융2팀, 계량팀

HW0

1. **아래 박스에서 모르는 단어들이 있다면, 관련 공부해오기**

Word Salad

벡터, 행렬, 미분, 적분, 최대값, 최소값, 중앙값, 기대값, 분산, 공분산, 확률분포, 확률변수, 누적분포함수, 상관관계, 상관계수, 인과관계, 독립, 종속, 정규분포, t분포, 카이제곱분포, F분포, IID, 추정, 검정, 함수, 전치행렬, 행렬식, 대각행렬, 직교, 내적, 분위수, 절대값, 확률밀도함수, 가설, 추정량, 신뢰구간, 제1종의 오류율, 제2종의 오류율, 검정력

2. 아래의 지시를 따라 필요한 프로그램 설치해오기
   $\Rightarrow$ 2시간 이상 고민하지 않은 오류에 대한 질문 안 받음
   $\Rightarrow$ Google, StackExchange, ChatGPT와 친해지자
3. (코딩 미경험자인 경우) 백준 기초 문제 다 풀고 정답 맞히기 (https://www.acmicpc.net/workbook/view/2196)

HW0 Instruction

1. Python 설치하기(https://wikidocs.net/186656)
2. Visual Studio Code 설치하기(https://wikidocs.net/187040)
3. VSCode에서 Jupyter, Python 관련 익스텐션 설치하기 (https://yeomss.tistory.com/227)
4. VSCode에서 Jupyter Notebook 관련 익스텐션 설치하기 (https://happygunja.tistory.com/47)
5. git 설치, GitHub 가입, GitHub Desktop 설치와 연동, 그리고 연습 (https://codegear.tistory.com/37)
6. 이 ppt에서 `package`처럼 쓰인 모든 패키지 다운로드 (https://dojang.io/mod/page/view.php?id=2443)
7. Overleaf 회원가입과 관련사이트 확인 (https://m.blog.naver.com/ljy5995/221680129935)[4]

[4] 이 글에서 LaTeX를 라텍스라 부르는데, 레이텍이 올바릅니다.

HW0 FAQ

- 파이썬 설치 관련(https://qucdas.tistory.com/210)
- 패키지가 설치가 안됨(https://statools.tistory.com/305)
- 깃헙 데스크탑 관련(https://turtledeveloper.tistory.com/27) (https://fishersheep.tistory.com/262)
- 깃헙 데스크탑 대신 git으로 직접 해도 되나요? ⇒ 됩니다[5]. (https://gin-girin-grim.tistory.com/10)
- 깃헙이랑 깃 연동이 안됩니다 (https://wikidocs.net/81070)
- 코드 틀린게 없는거 같은데 안돌아가요 ⇒ 그대로 복붙해서 구글검색 (https://blockdmask.tistory.com/550)
- 파이썬이 파일을 못 찾을 경우 (https://gr-st-dev.tistory.com/1891)
- 다른 IDE 쓰고싶어요 ⇒ 됩니다[6].

[5] 그치만 이미 git 사용할 줄 아시는 분은 이걸 안 들으실듯

[6] 특히 경통 등의 수업에서 Jupyter Lab 사용하신 분들

Etc

- 질문: 이론 및 새로운 것에 대한 질문은 적극 환영합니다! 하지만 프로그램 설치 등에 대한 질문은 저도 다시는 하기 싫은 고통스러운 과정인터라, 최대한 받지 않도록 하겠습니다(구글을 적극 활용하세요)
- 주의: 저도 파이썬으로는 잘 안해봐서 서투릅니다.
- 주의: 일정은 자료 만들면서 바뀔수 있습니다.
- 어느정도 수준?: 내용들은 통계/수리/컴공/데싸 학부2학년~대학원 수준까지 다양하지만, 간단하게 응용 위주로 다루는지라 경수/경통/컴개실 수준만 있어도 무리없을 것이라 기대합니다. 계량/수통/경통연을 들으신 경우 쉽게 이해하실 수 있는 수준입니다.
- 로드: 강의만 들으면 실력이 늘기는 쉽지 않아서, 숙제 열심히 하시고 이후에도 열심히 하실 분들이 수강하시면 좋겠습니다. 열심히 하실 경우 로드는 실습 낀 1~2학점 SU과목 수준으로 예상합니다.