데이터 분석을 위한 참고자료 *

신수안

January 14, 2020

1 논문 작성을 위한 가이드

- **King (2006)**: a short article about "how to write a publishable paper by beginning with the replication of a published article". 마지막 "Problems to Avoid and Other Suggestions (p. 124)"이 특히 유용하다.
- **Prof. Kosuke Imai의 노트**: 대학원 방법론 수업의 final project에 대한 안내문. 마지막 "General Tips for Writing a Scientific Paper"에서 논문을 작성하는 구체적인 단계에 대해 조언하고 있다.
- **Abstract 작성법**: how to construct a *Nature* summary paragraph.

2 기본기 쌓기

- **Harvard Gov Prefresher**: Dep. of Government 신입생들에게 입학 전 제공되는 math camp의 오픈 소스 데이터. 수학과 R 프로그래밍이 어느 정도 익숙한 사람도 Ch. 13-15의 LaTeX, regular expression, command-line, git에 대한 간략한 소개는 읽어보면 도움이 될 것이다. pdf도 다운받을 수 있다.
- **QSS 코드**: Imai (2018)의 데이터와 코드가 포함된 R package. 사회과학에 초점을 맞춘 좋은 입문서. 여기에 다른 참고자료에 대한 안내도 나와있다.
- 그밖의 **R tutorials**

3 데이터 찾기

- **ICPSR**: 미시건 대학에서 제공하는 데이터베이스 (제휴기관에 한해 다운로드 가능). 동명의 summer program이 존재한다.
- **Harvard Dataverse**: replication data를 모아둔 데이터베이스 (오픈소스).

*해당 자료는 2020 서울대학교 정치외교학부 방법론캠프 수업자료의 일환으로 작성되었으며, 작성자의 지극히 개인적인 경험을 바탕으로 하였음을 밝힙니다.

- **World Economics and Politics Dataverse**: comparative/international political economy에 관한 데이터베이스.
- **KOSSDA**: 한국사회과학자료원의 데이터베이스. KOSSDA에서는 주기적으로 방법론 단기강좌를 열고 있기도 하다.
- 이밖에도 Comparative Agendas Project, Voteview, ANES, WVS, Lobbyview 등 다양한 (검색 기반의) 데이터베이스가 존재하며 미국 주요대학의 도서관 웹사이트, 연구자의 홈페이지 등에서도 리스트를 찾을 수 있다.

4 데이터 분석하기

4.1 Preprocessing

- **Tidyverse**: Hadley Wickham은 R 코딩의 새로운 트렌드를 주도했고, 그 중심에 tidyverse가 있다. Tidyverse는 데이터 분석에 유용한 R package들을 묶어놓은 하나의 컬렉션이다. 주로 readr, dplyr, tidyr, purrr, ggplot2의 function들이 유용하게 쓰인다. Tidyverse에 적합한 코딩 스타일은 여기서 찾을 수 있으며, 특히 pipes에 관한 내용을 눈여겨볼 만하다. 참고로 tidyverse를 구글 이미지 검색하면 패키지들 사이의 관계를 쉽게 파악할 수 있을 만한 그림들을 찾을 수 있다.
- **Wickham and Grolemund (2016)**: tidyverse를 이용한 데이터 분석의 좋은 참고서. Ch. 5와 12가 특히 유용하다.
- **Silge and Robinson (2017)**: tidy format을 활용한 텍스트 분석 개념서. 텍스트 자료를 분석하려 하는데 어디서부터 어떻게 시작해야 할지 막막할때 참고할만하다. word embedding에 대한 내용은 없다는 것이 작은 흠.

4.2 Visualization

- **Wilke (2019)**: 내가 가진 데이터를 어떻게 하면 가장 효과적으로 visualize 할 수 있을지 아이디어를 얻기 좋은 참고서. 심미적인 부분에 대해 굉장히 세세하게 조언하는 책이다. 이 책을 만드는데 쓰인 코드는 여기서 찾을 수 있다.
- **Practical ggplot2**: 위 책에서 소개된 주요 plot들에 대한 상세한 코드를 확인할 수 있다.

4.3 Advanced Programming

- **Wickham (2014)**: R programming에 대한 깊은 이해를 위한 참고서. Ch. 12-16의 OOP system과 데이터 타입(S3, S4)에 관한 내용, Ch. 25의 Rcpp에 관한 내용이 특히 유용하다.
- **Wickham (2015)**: R package의 구성요소들을 파악하거나 직접 패키지를 만드려 하는 이들을 위한 참고서. Methodologist의 경우 자신이 개발한 method의 코드를 R package로 만들고 (GitHub repository로 공개하거나) CRAN에 등록한다.

5 유용한 플랫폼

- **Overleaf**: 온라인 LaTeX편집기.
- **GitHub**: 대표적인 Git GUI.
- **Slack**: cowork에 유용한 메신저.

6 코드를 쓰다가 헷갈릴 때

- R Cheat Sheets: R studio 사이트에서 쉽게 찾을 수 있다.
- LaTeXCheat Sheets: 자료1, 자료2, 자료3 등.

7 방법론을 더 공부하고 싶을 때

주변에서 관심있는 주제의 강의를 찾기 힘들다면, 교수들의 웹사이트에 공개된 실라버스 및 강의 자료를 참고하자. 미국의 몇몇 정치학과들은 3-4개의 Methods sequence가 존재하며, 이들 실라버스를 살펴보아 관심 있는 method의 reading materials를 집중적으로 공부해볼 수도 있을 것이다.
E.g. Prof. Kosuke Imai (Harvard), Prof. Matthew Blackwell (Harvard), Prof. Teppei Yamamoto (MIT), Prof. In Song Kim (MIT), Prof. Yuki Shiraito (U. Michigan).

References

Imai, Kosuke. 2018. *Quantitative Social Science: An Introduction*. Princeton University Press.

King, Gary. 2006. "Publication, Publication." *PS: Political Science & Politics* 39(1):119–125.

Silge, J. and D. Robinson. 2017. *Text Mining with R: A Tidy Approach*. O'Reilly Media.

Wickham, H. 2014. *Advanced R*. Chapman & Hall/CRC The R Series CRC Press.

Wickham, H. 2015. *R Packages: Organize, Test, Document, and Share Your Code*. O'Reilly Media.

Wickham, H. and G. Grolemund. 2016. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.

Wilke, C.O. 2019. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. O'Reilly Media.