POLS 537: Advanced Research Methods and Data Analysis in Political Science

Instructor: Dr. Serkant Adiguzel (serkant.adiguzel@sabanciuniv.edu)
Office Hours: Mondays 2:00 pm-4 pm (FASS 2093)

Course Schedule: Mondays: 11:40 am - 1:30 pm (FASS 1080).
Tuesdays: 12:40 pm – 1:30 pm (FASS 1099).

Course Description and Objectives:

Social scientists use various quantitative methods using different types of quantitative administrative, media, social media, spatial, and audio data. This course will introduce you to some of these methods and give you the ability to understand and perform such research independently.

The course will emphasize quantitative data analysis and improve your data science skills. Over the last two decades, social science research that uses quantitative data has flourished. Similarly, many organizations, such as NGOs, corporations, and governments, use data to make informed decisions and need people with the necessary data skills. Therefore, obtaining such skills is valuable beyond academia.

The course aims to maintain a balance between theory and application. Therefore, the lectures will include various applications and live-coding sessions, and you will also be expected to apply what you learn with problem sets and the data project.

The course is designed in five parts. The first part will focus on everyday data analysis problems, which are usually not taught in methods classes but learned as you do your research. The first two weeks will focus on data types and cleaning/processing/discovering/presenting your data. It will also focus on automated data collection methods with APIs and web scraping.

The following three parts focus on estimation, prediction, and discovery. In the second part (estimation), we study causal inference with experimental and observational data. The third part will focus on prediction using linear regression and its variants (LASSO and Ridge) and predictive modeling in general. Part 4 will focus on discovery in three domains: network, spatial, and text data.

The last part will specifically focus on text data, covering unsupervised and supervised models using text data.

Prerequisites:

Students are expected to take POLS 530 Quantitative Research Methods before taking this course.

We will use R (and sometimes Python) in this course for data analysis. R is a free and open-source programming language used by data scientists, mainly for data analysis and visualization. RStudio is an integrated development environment (IDE) for R. You should install both R (https://www.r-project.org) and RStudio (https://www.rstudio.com/products/rstudio/download/) on your computers. Python, on the other hand, is a high-level general purpose programming language.

Although not required, you can install Python by following steps in this tutorial: https://docs.python-guide.org/starting/installation/. Any python script will be distributed as a notebook on Google Colab so you do not have to install Python.

Although R is a simple and intuitive programming language, it can initially have a steep learning curve. Therefore, I will use office hours and the first three weeks in class to help you learn it. I will also hold live coding sessions during lectures so I strongly suggest that you bring your laptops so that you can code with me.

If you never used R before, I strongly suggest that you become acquainted with the basic syntax before coming to the first class. Please complete this online tutorial if you never used R before: (https://campus.datacamp.com/courses/free-introduction-to-r/).

Student responsibilities:

To receive a passing grade from the course, you **must** complete all assignments –i.e., all problem sets, reaction points/quizzes, replication/extension exercise, and the data project. The distribution of the grade is going to be as follows:

Reaction points/quizzes (20 percent): Starting with Week 3, you are expected to do weekly assigned readings before coming into the class. The required readings will be available on SUCourse. To ensure that we are all keeping up, please post a half page or so of discussion/reaction points and/or questions bearing on the week's reading to the SUCourse by 8 pm before class (i.e., by Sunday 8 pm). You can raise questions about the methodologies used in papers, question their assumptions, the validity of their results, etc. You can also highlight any points you find confusing or did not understand! Or you can compare readings within the same week or develop links across weeks regarding methodologies pursued. It is up to you! All I require is that the points need to be thoughtful, and they do not need to be long. If you make some trivial comments (such as: "this is a very interesting paper, and it is about x and y") that clearly show that you did not do the readings, you will not get any points for that week. I will also do pop-up quizzes about the assigned readings. These reaction points and quizzes will constitute 20% of your total grade.

Student attendance/participation (10 percent): You are expected to attend lectures and participate discussions actively. Therefore, attendance will constitute 10% of your overall grade.

Replication/extension exercise (10 percent): Each student is expected to conduct a replication and extension analysis using quantitative data from an academic paper. Replications/extensions should replicate a paper's results, check the robustness of the original results (by, for instance, analyzing subsamples, testing model assumptions, estimating different models, etc.), and engage in at least one extension that is theoretically or substantively important. The extension could extend the analysis to a different sample, time or include a separate analysis, such as a heterogeneity analysis. The extension can also be about testing a different yet alternative hypothesis about the paper analyzed. You should use at least one method we learn during class in these extensions. You will provide a 20-minute presentation and a 2-page report based on your analysis. The deadline for reports is **one week** after your presentation. Hence if you do your presentation on, say, Week 5, the deadline for the report is Week 6 before class (11:40 am)!

We will start replication/extension exercises in Week 5, and each student will sign up for a particular week. The paper chosen for the exercise should be relevant to the course's theme in terms of the methods employed, and its data should be publicly available. Before starting the analysis, you will need my approval for the chosen paper since I need to confirm that you can use one of the methods we learn and that its data/code is available. I will send a sign-up sheet after Week 1 where you can sign up for a particular week on a first come first served basis.

Data project (20 percent): You will need to specify a clear research question that can be analyzed with quantitative data (administrative data, social media data, survey data, etc.). You will be responsible in finding the data. You will analyze the data to answer your research question and write no more than a 10-page report that will be due **TBD**. You will have to write a 2-page project proposal in which you talk about your research question and the dataset you will use by **April 8, 2024, at 8 pm** so that I can give you detailed feedback about your project. You will also have to arrange at least **two** meetings with me to talk about your progress (one before April 8 and the other after April 8) (you can sign up for office hours using the link).

You can use this data project as an opportunity to make progress in your MA/Ph.D. theses. Hence, you are encouraged to submit the quantitative data analysis part of your theses as the data project for this course. Use this as an opportunity to make progress in your research! I will also organize an informal workshop where you will have a chance to present your projects in class. This will take place at the end of the semester (time and location: **TBD**). The deadline for the data project will be **after** the workshop so that you can incorporate the feedback into the final version.

Graded problem sets (40 percent): There will be **five** graded problem sets during the semester. Using real-world data, you will be asked to apply the things you learn during the lectures. You can work in groups (2 people max) for these problem sets but everyone should submit their own answers and code. Therefore, you will be expected to submit your own solutions, but you should write your collaborators' names for each submission. You will use RMarkdown in these problem sets since I will grade your code and solutions together. Learn how to use RMarkdown and come to the office hours whenever you have questions.

Grade distribution:

Reaction points/quizzes: 20%
Student attendance/participation: 10%
Replication/extension exercise: 10%
Data project: 20%
Graded problem sets: 40% (8% each)

Class Policies and Rules:

- There is no margin or font requirement for written assignments. However, they need to be professional-looking! This means that it should include page numbers, proper citations and formatted bibliography, formatted tables and/or figures. I suggest you use RMarkdown since it will allow you to easily insert your results into nicely formatted tables or figures.

- You are required to submit your work on time. Late submissions will be penalized by 2 points for each hour they are late. Late submissions for reaction points will not be accepted.

- In line with the university's academic integrity statement, you are expected to base your work on your labor and ideas in this class. Therefore, plagiarism will not be tolerated and will result in a letter grade F and further disciplinary action.

- If you would like to come to my office hours, please use this google sheet before coming: https://bit.ly/3LrphQa. This way, you won't have to wait for your friend in front of the office. Depending on your question, you can sign up for as many 20-minute slots as possible. Please email me ONLY when you cannot sign up for any time slot due to unavailability or your schedule.

Textbooks:

There is no required textbook for this course since all assigned readings will be on SUCourse. However, the following textbooks are useful reference books.

1) Baruffa, O. & others (2023). *Big Book of R*. This is an open-source collaboration that is a book of books. Available here: https://www.bigbookofr.com/index.html

2) Imai, K., & Williams, N. W. (2022). *Quantitative Social Science: An Introduction in Tidyverse*. Princeton University Press (IW, henceforth).

3) Diez, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. 2015. *Open-Intro Statistics*. 3rd edition (https://www.openintro.org/book/os/)

4) Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press (IR, henceforth).

5) Tunstall, L., Von Werra, L., & Wolf, T. (2022). *Natural language processing with transformers*. O'Reilly (TVW, henceforth)

6) Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press (GRS, henceforth).

Course Overview

WEEK 1 (19-20 February 2024): Introduction to course logistics and R & Everyday problems during research and data analysis (1):

- Data types in R and beyond
- Data cleaning/processing/discovering your data
- Data visualization and presentation (ggplot2 and Shiny)

WEEK 2 (26-27 February 2024): Everyday problems during research and data analysis (2): Data collection with APIs and web scraping

⇒ First Problem Set is available

- Using APIs in R and Python
- Web scraping with R libraries (static websites)
- Dynamic web scraping with Python (selenium)

Readings:

- Bauer, P.C., Landesvatter, C., Behrens, L. (2021). APIs for social scientists: A collaborative review, Available here: https://bookdown.org/paul/apis_for_social_scientists/ (Read Chapter 1, 2 and skim the rest)

WEEK 3 (4-5 March 2024): Causal inference with experiments (1)

- Potential outcomes framework
- Problem of causal inference
- RCT's
- Difference-in-means and regression

Readings:

- Textbooks: IR Chapters 1, 2, 3, 4, 7; IW Chapter 2.4, 4.3
- Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 73(1), 31-43.
- Angrist, J. D., & Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2), 3-30
- Teele, D. L. (Ed.). (2014). *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*. Yale University Press. (Chapters 1, 2)
- Gerber, A. S., Green, D. P., & Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, 102(1), 33-48.

WEEK 4 (11-12 March 2024): Causal inference with experiments (2)

⇒ First Problem Set is due

⇒ Second Problem Set is available

- Possible problems: Noncompliance, attrition, poor data quality, data collection
- Pre-analysis plan
- p-hacking and ethical considerations

Readings:

- Textbooks: IW Chapters 3.1, 3.2, 7.2.5
- Teele, D. L. (Ed.). (2014). *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*. Yale University Press. (Chapters 7, 9)
- Franco, A., Malhotra, N., & Simonovits, G. (2015). Underreporting in political science survey experiments: Comparing questionnaires to published results. *Political Analysis*, 23(2), 306-312.
- Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1), 1-32.

WEEK 5 (18-19 March 2024): Causal inference with observational data (1)

- Difference-in-differences

Readings:

- Bookmark this: https://asjadnaqvi.github.io/DiD/

- David Card and Alan Krueger (1994) "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania." *American Economic Review*, 84(4), 772–793.
- Roth, J., Sant'Anna, P. H., Bilinski, A., & Poe, J. (2022). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. arXiv preprint arXiv:2201.01194.
- Baker, A. C., Larcker, D. F., & Wang, C. C. (2022). How much should we trust staggered difference-in-differences estimates?. *Journal of Financial Economics*, 144(2), 370-395.

WEEK 6 (25-26 March 2024): Causal inference with observational data (2)

- Sharp and fuzzy regression discontinuity designs (RDD)

Readings:

- Textbook: IW Chapter 4.4.3
- Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, 48(2), 281-355.
- Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, 37(3), 447-456.
- Eggers, A. C., & Hainmueller, J. (2009). MPs for sale? Returns to office in postwar British politics. *American Political Science Review*, 103(4), 513-533.
- Dell, M. (2010). The persistent effects of Peru's mining mita. *Econometrica*, 78(6), 1863-1903.

- Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6), 2295-2326.

WEEK 7 (1-2 April 2024): Prediction

⇒ PS 2 is due.
⇒ PS 3 is available..

- OLS revisited
- Model Fit
- Out-of-sample performance and overfitting
- LASSO, Ridge Regression, and Elastic Net

Readings:

- Cranmer, S. J., & Desmarais, B. A. (2017). What can we learn from predictive modeling?. *Political Analysis*, 25(2), 145-166.
- Hindman, M. (2015). Building better models: Prediction, replication, and machine learning in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, 659(1), 48-62.
- Mueller, H., & Rauh, C. (2018). Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*, 112(2), 358-375.

8-9 April 2024: NO CLASS (SPRING BREAK)

WEEK 8 (15-16 April 2024): NO CLASS

WEEK 9 (22 April 2024): Discovery

⇒ PS 3 is due.
⇒ PS 4 is available.

- Discovery with network data
- Discovery with spatial data

Readings:

- Textbook: IW Chapter 5.
- Padgett, J. F., & Ansell, C. K. (1993). Robust Action and the Rise of the Medici, 1400-1434. *American Journal of Sociology*, 98(6), 1259-1319.
- Fernando, G. A., & Antoine, M. (2022). The network structure of global tax evasion evidence from the Panama papers. *Journal of Economic Behavior & Organization*, 197, 660-684.

WEEK 10 (29-30 April 2024): Text as data (1)

- Bag of words
- Word embeddings

Readings:

- Textbook: GRS, Chapters 5 and 8.
- Rodriguez, P. L., & Spirling, A. (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics*, 84(1), 101-115.
- Rodriguez, P. L., Spirling, A., & Stewart, B. M. (2021). Embedding Regression: Models for Context-Specific Description and Inference. *American Political Science Review*, 1-20.
- Almelhem, A., Iyigun, M., Kennedy, A., & Rubin, J. (2023). Enlightenment Ideals and Belief in Progress in the Run-up to the Industrial Revolution: A Textual Analysis. *Working Paper*

WEEK 11 (6-7 May 2024): Text as data (2): Unsupervised models

- Clustering
- Topic models

Readings:

- Textbook: GRS, Chapters 12 and 13.
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4), 1064-1082.
- Greene, D., & Cross, J. P. (2017). Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, 25(1), 77-94.
- Bianchi, F., Terragni, S., & Hovy, D. (2020). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint* arXiv:2004.03974

WEEK 12 (13-14 May 2024): Text as data: Supervised models and text classification

⇒ PS 4 is due.
⇒ PS 5 is available.

- Naïve Bayes
- Random forests
- Neural networks
- Transformer-based large language models

Readings:

- Textbook: GRS, Chapters 17, 18, 19
- Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, 29(1), 19-42.
- Jacobs, A. M., Matthews, J. S., Hicks, T., & Merkley, E. (2021). Whose news? Class-biased economic reporting in the United States. *American Political Science Review*, 115(3), 1016-1033.
- Licht, H. (2022). Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings. *Political Analysis*, 1-14.

WEEK 13 (20-21 May 2024): Text as data: Supervised models and text classification (2)

Readings:

- Textbook: TVW, Chapters 1, 2
- Terechshenko, Z., Linder, F., Padmakumar, V., Liu, M., Nagler, J., Tucker, J. A., & Bonneau, R. (2020). A comparison of methods in political science text classification: Transfer learning language models for politics. Available at SSRN 3724644.
- Widmann, T., & Wich, M. (2022). Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in german political text. *Political Analysis*, 1-16.
- Wankmüller, S. (2021). Introduction to Neural Transfer Learning With Transformers for Social Science Text Analysis. *Sociological Methods & Research*, 00491241221134527.

WEEK 14 (27-28 May 2024): Large language models and beyond

⇒ PS 5 is due.

Readings:

- Check this! https://bbycroft.net/llm
- How do transformers work? https://osanseviero.github.io/hackerllama/blog/posts/random_transformer/
- Ludwig, J., & Mullainathan, S. (2024). Machine Learning as a Tool for Hypothesis Generation. *Quarterly Journal of Economics*