Research Methods I

Data Analysis for Economics and Policy

Instructor Information

- **Instructor**: Fernando Rios-Avila
- **Email**: friosavi@levy.org

- **Office Hours**: Wednesdays 1:30pm to 4:00pm. Or by appointment. Other times can be arranged, but will be done remotely.
- **Course Website**:
- **Class Time**: Wednesday, 9:30 am - 12:45 am

1 Course Description

This course focuses on providing students with the tools and skills necessary to conduct data analysis for economics and policy research. Students will be exposed to the entire process of data analysis, from formulating questions and collecting data to cleaning, exploring, analyzing, and presenting results. The course covers exploratory data analysis, regression analysis, and introduces topics on prediction with machine learning. Students will gain hands-on experience using Stata, with Quarto for reproducible reporting, and GitHub for version control and collaboration.

2 Course Objectives

By the end of this course, students will be able to:

1. Apply advanced data analysis techniques to economic and policy questions.
2. Use modern tools such as GitHub and Quarto for research collaboration and reproducibility.
3. Formulate research questions and design appropriate data collection methods.
4. Clean, organize, and explore data using various techniques and visualizations.
5. Apply regression analysis techniques to analyze relationships between variables.
6. Use machine learning methods for prediction and classification tasks.
7. Implement data analysis techniques using Stata.
8. Effectively communicate research findings through written reports and oral presentations.

3 Required Textbook

Békés, G., & Kézdi, G. (2021). Data Analysis for Business, Economics, and Policy. Cambridge University Press.

4 Software Requirements

- `Stata`: A student license will be provided.
- `Quarto`: Free and open-source software for reproducible research.
- `VSCode`: Free and open-source code editor.
- `GitHub/GitHub-Desktop`: Free platform for version control and collaboration.
- `zotero`: Free reference manager.

> **!** Important
>
> All homework assignments are required to be submitted in Quarto format, using GitHub to submit the assignments.

5 Course Outline

1: Introduction to Modern Research Tools

- Course overview and expectations.
- Introduction to GitHub/Github-Desktop for version control and collaboration.
- Getting started with Quarto for reproducible research: RStudio and VSCode.
- Other Tools: Overleaf, Zotero
- Data organization and management
- **Lab**: Setting up GitHub, Quarto, VSCode, Zotero, and Overleaf

2. Introduction to Data Analysis

- Introduction to data analysis: The Process
- What is Data? What types of Data are there?
- Data collection methods
- Preparing data for analysis
- Tidy data principles
- Data cleaning: Missing values, outliers, and errors
- **Readings**: Chapter 1 and 2

3: Data Exploration

- Type of data vs type of analysis
- Frequencies, distributions, and summary statistics
- Exploratory data analysis techniques and visualizations
- Theoretical distributions
- Comparisons, correlations and conditional distributions
- Latent and observed variables
- **Readings**: Chapter 3 and 4

4: Generalization: From Sample to Population

- Sampling and generalization
- Repetition and sampling variability
- Confidence intervals and standard errors: The Bootstrap method
- External validity
- Hypothesis testing principles
- Type I and Type II errors
- Multiple Hypothesis Testing and p-hacking
- **Readings**: Chapters 5 and 6

5 and 6: Regression Analysis I: Simple Regression

- Linear and non-linear relationships
- Linear Regression: Estimation and interpretation
- Correlations and coefficients: Searching for causality
- Properties and assumptions of the linear regression model
- Transformations and Semiparametric models
- Extreme values, Influential observations and measurement error
- Generalizing Results: SE and CI
- Testing Hypotheses
- **Readings**: Chapters 7, 8, and 9

7: Regression Analysis II: Multiple Regression

- Multiple regression basics: Estimation and Inference
- Problems with multiple regression
- Non-linearities, interactions and qualitative variables
- **Readings**: Chapters 10

8: Regression Analysis III: Modeling Probabilities

- Linear Probability Model
- Non-linear models: Logit and Probit
- Interpretation and marginal effects
- Goodness of Fit and Predictive Power
- **Readings**: Chapter 11

9: Time Series Analysis

- Introduction to time series data
- Trend and seasonality
- Stationarity and autocorrelation
- Serial correlation
- **Readings**: Chapter 12

10: Prediction

- Introduction to Prediction
- R2 vs AR2 and other measures of fit
- Overfitting and Cross-validation: Finding the right model
- External validity and generalization
- **Readings**: Chapter 13

11: Model Building for Prediction: LASSO

- The Process of Prediction
- How to choose $g(y)$
- Working with $X's$
- Introduction to LASSO: Prediction and Diagnosis
- **Readings**: Chapter 14

12: Predicting Probabilities and Classification

- Predicting Probabilities and Classification
- Classification, confusion matrices, and ROC curves
- Finding the right threshold
- **Readings**: Chapter 17

13: Forecasting Data

- Introduction to Forecasting: Predicting the future

- Training and Testing Data
- Trends, Seasonality, and Cycles
- Forecasting with ARIMA
- VAR and External validity
- **Readings**: Chapter 18

6 Grading Policy

- **Weekly Quizzes**: 10%
  - Short quiz at the end of each class to test material covered.
- **Weekly Problem Sets**: 30%
  - Small problem-sets of one or two questions to be completed and submitted in Quarto format using GitHub. This may include, data exploration and visualization, regressions and interpretation, and prediction. It also includes the peer review of another student's work.
- **Term Paper**: 60%
  - Multi-part research project applying techniques learned in class to real data.

Term Paper (60% of final grade)

Throughout the semester, students will work on a multi-part research project that applies the techniques learned in class to real-world data. This project will require students to propose a research question, collect and clean data, conduct exploratory and regression analyses, and present their findings.

Part I: Research Proposal (5%, due Week 2)

- Introduction (including research question and motivation)
- Proposed data sources: Consider data from the textbook, Kaggle, or other sources
- Expected findings and relevance
- Conclusion
- References (if any)
- Specify which software (if other than Stata) you plan to use for your analysis
- Create a GitHub repository for your project and submit the link with your proposal

Part II: Data Collection and Cleaning (10%, due Week 4)

Write a report that includes the following:

- Description of data sources, collection methods, and potential biases
- Data overview and preprocessing steps

- Discussion of issues encountered and solutions
- Preliminary analysis with descriptive statistics and visualizations
- Include a data dictionary in your GitHub repository

Peer Review Report (due Week 5)

- Review another student's work and provide feedback on their data collection and cleaning process. Provide suggestions for improvement and identify any potential issues.

Interim Progress Report (5%, due Week 7)

- Brief update on progress, challenges faced, and next steps
- Additional visualizations or analyses
- It should include a draft of literature review and methodology
  - For the literature review include summaries for 3-5 papers related to your research question.
  - It should also include a draft of the methodology section, including the model specification.
- Address any feedback received from the peer review in Part II

Part III: Data Analysis (15%, due Week 10)

Present a draft for data analysis that includes the following:

- Model specification discussion
- Regression results presentation
- Interpretation of results
- Discuss Limitations and robustness checks

Peer Review Report (due week 11)

- Review another student's work and provide feedback on their data analysis. Provide suggestions for improvement and identify any potential issues.

Part IV: Final Report (20%, due Week 13)

Complete research paper should include:

- Introduction
- Literature Review
- Data and Methodology
- Robustness Checks or Sensitivity/Sub-group Analysis
- Conclusion

- References
- Appendices (if any)

Presentaton (5%, Week 14)

- 15-minute presentation of your research to the class
- See here for an example for the kind of report expected at each stage, based on the first research proposal on the impact of remote work on urban housing prices.

Additional Requirements

- All work should be submitted in Quarto format using GitHub.
- Unless Data used is confidential or too large, you should include all data in your GitHub repository.
- Your GitHub repository should include with all code, data (if possible), and the Quarto document for your report.
- It should also include all papers used for the literature review.
- At each stage, you should submit an email with the PDF and QMD files to the instructor, at or before the deadline. You should also *push* your work to GitHub to follow the progress of your project.

Resources

- **Textbook**: Békés, G., & Kézdi, G. (2021). Data Analysis for Business, Economics, and Policy. Cambridge University Press. There are additional resources available on the book's website: Data Analysis
- **Software**: There are several statistical packages for analyzing data. In this course, we will be using the software `Stata` to cover all materials in class. Slides are self-replicable, thus you can copy and paste almost all code provided to replicate the results seen in class. The Institute will be providing you with licenses for Stata/BE for the length of the course.

Stata offers many free short webinars and video tutorials that may be useful if you never used Stata before, or even if you have some experience with it. Please see the resources page for more information.

If you decide to, you can also use R, Julia, or Python to study and work on the course materials and homework. The book for the class has a repository with all the code in Stata, R and Python. It could be of great advantage to you to learn other languages, as they are widely used in the industry and academia.

As with many other skills, the best way to learn is to simply work with the software, work on the book exercises, and ask any questions to me or your classmates when you find a problem you could

not find a solution for.

For the additional software, please look into Quarto, GitHub, Zotero and VSCode.

7 Course Policies

- **Attendance**: Attendance is highly recommended. Classess will not be recorded, and except for exceptional cases, there will be no online classes.
- **Late Assignments**: Late assignments will not be accepted unless prior arrangements have been made with the instructor.
- **Academic Integrity**: All work submitted must be your own. Plagiarism will not be tolerated and will result in a failing grade for the assignment or course.
- **AI usage**: The use of AI in the class is allowed. However, you must disclose any AI tools used in your assignments. AI is a tool you can use to generate ideas, edit your text, provide help with coding, etc. However, it is completely unacceptable to use AI to generate the entire assignment. You will have to be able to explain and defend your work in class.