

APPLIED DATA ANALYSIS

Faculty of Economics and Finance

Department of Economic Policy

Instructor:	Eduard Nezinsky, Tomas Oles, Ella Sargsyan
Email:	eduard.nezinsky@euba.sk; tomas.oles@euba.sk; esargsyan@cerge-ei-foundation.org
Office Hours:	Thursday: 1-3:00 PM
Office Location:	Old Building, 4B34; 4B40
Credit Hours:	6
Term:	Winter 2024

COURSE DESCRIPTION

The first goal of the course is to teach students knowledge and skills in modern methods of applied data analysis and statistical learning, including the use of R software, to conduct empirical economic research and design research approaches and methods for solving economic problems.

The second aim of this course is to acquaint students with basic econometric concepts and methods which are necessary to read and understand current empirical research in economics, as well as plan and execute independent research projects. The course will introduce the idea of regression analysis and cover some of the most recent econometric techniques central to modern econometric practice, which aims to identify causal relationships.

STUDENT LEARNING OUTCOMES

Upon completing the course, students should acquire knowledge of modern methods of data research and visualization, classification, linear regression, and general data analysis. Skills in working with data, which they will be able to apply in their own empirical research. They will also gain advanced skills in using modern software (R) in empirical economic research and will be able to write their own functions.

Students will know basic econometric concepts, basic estimation methods, and methods for testing statistical hypotheses. They will be able to apply standard methods of constructing econometric models, process statistical information, obtain statistically sound conclusions, and give meaningful interpretation to the results of the estimated econometric models. In addition, students will gain real data processing skills, using econometric packages for building and estimating regression models in R.

PREREQUISITES

Statistics.

TEXTBOOKS:

Wickham, H., Çetinkaya-Rundel, M., Grolemund, G. (2023). R for Data Science. O'Reilly Media, Inc.. The book is available for free at <https://r4ds.had.co.nz/>.

Gareth, J., Daniela, W., Trevor, H., Robert, T. (2013). An Introduction to Statistical Learning: With Applications in R. Springer. The book is available for free at <https://www.statlearning.com/>.

Wooldridge, Jeffrey M. Introductory Econometrics: A Modern Approach. 5th ed., US, 2012

Angrist, J. D., Pischke, J. S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

"We are what we repeatedly do. Excellence, then, is not an act, but a habit." Aristotle

GRADING AND EVALUATION

The course evaluation is based on **two components**, which together determine the final grade. To complete the first component that is condition for obtaining the CERGE-EI certificate (60% of the final grade):

1. Quizzes and Assignments (20%):

- Throughout the course, students will receive quizzes or short assignments during lectures. These will account for 20% of the final grade.
- Active participation and extra involvement in these quizzes or assignments may also positively influence this portion of the grade at the instructor's discretion.

2. Home Assignments (30%):

- There will be four home assignments, each consisting of a small data science project. Examples of these projects include:
 - Mapping COVID-19 incidence across different regions.
 - Calculating the wage premium for workers in the U.S.

3. Final Exam (50%):

- The final exam will consist of:
 - A *theoretical section*, which will cover the properties, derivation, and assumptions and interpretation of linear and logistic regression, as well as basic causal research designs or panel data techniques.

The second component to obtain a full number of points from the course (40% of the final grade):

1. Quizzes and Assignments (50%):

- Throughout the course, students will receive quizzes or short assignments during and after lectures. These will account for 50% of the final grade.
- Active participation and extra involvement in these quizzes or assignments may also positively influence this portion of the grade at the instructor's discretion.

2. Final Data Science Project (50%):

- The final presentation of data science project.

Final Grade Calculation: Students do not need to achieve at least 50% in each individual component. The final grade will be based on the cumulative score across all components.

IMPORTANT DATES

The instructor reserves the right to change the content of the course material if he perceives a need due to postponement of class caused by inclement weather, instructor illness, etc., or due to the pace of the course.

FINAL EXAM: December 9 – 13, 2024

MAKEUP EXAM: December 16 – 20, 2024

ACADEMIC INTEGRITY:

Academic integrity is essential to the pursuit of learning and scholarship in a university, and to ensuring that a degree from the University of Economics in Bratislava is a strong signal of each student's individual academic achievement. As a result, the University treats cases of cheating and plagiarism very seriously. The University of Economics in Bratislava's Ethical Code (<https://euba.sk/univerzita/eticky-kodex>) outlines the behaviors that constitute academic dishonesty and the processes for addressing academic offenses. Potential offenses include, but are not limited to:

In papers and assignments:

Using someone else's ideas or words without appropriate acknowledgment; Submitting your own work in more than

one course without the permission of the instructor;
Making up sources or facts;
Obtaining or providing unauthorized assistance on any assignment.

On tests and exams:

Using or possessing unauthorized aids;
Looking at someone else's answers during an exam or test;
Misrepresenting your identity/pretending to be someone else.

In academic work:

Falsifying documents or grades;
Falsifying or altering any documentation required by the University, including (but not limited to) medical forms.

All suspected cases of academic dishonesty will be investigated following procedures outlined in the Ethical Code. If students have questions or concerns about what constitutes appropriate academic behavior or appropriate research and citation methods, they are expected to seek out additional information on academic integrity from their instructors or from other institutional resources.

CLASS SCHEDULE

Week	Topic Description	Day
1	<p>R basics: Basic R syntax, data types, vectors arithmetic, indexing, sorting, sorting using <code>dplyr</code>, and plotting using basic packages. GitHub.</p> <p><i>Reading</i> Gareth et al. (2013) An Introduction to Statistical Learning: With Applications in R. Chapter 1</p> <p>Wickham et al. (2023) R for Data Science Chapters 1, 2, 4-6</p>	Tuesday - lecture and exercise according to schedule in AIS according to schedule in AIS
2	<p>Visualization Part 1: Data visualization principles, creating custom plots with <code>ggplot2</code>, and studying the advantages and pitfalls of widely-used plots.</p> <p><i>Reading</i> Wickham et al. (2023) R for Data Science Chapters 3, 7</p>	Tuesday - lecture and exercise according to schedule in AIS
3	<p>Visualization Part 2: Data visualization principles, creating custom plots with <code>ggplot2</code>, and studying the advantages and pitfalls of widely-used plots.</p> <p><i>Reading</i> Wickham et al. (2023) R for Data Science Chapters 3, 7</p>	Tuesday - lecture and exercise according to schedule in AIS
4	<p>Data Summarizing, Tiding and Joining: Importing data from different file formats, web scraping, tidy data with <code>tidyverse</code>, wrangling data with <code>dplyr</code>, handling date and time formats.</p> <p><i>Reading</i> Wickham et al. (2023) R for Data Science Chapters 9-16</p>	Tuesday - lecture and exercise according to schedule in AIS
5	<p>Introduction to statistics and probability: Elementary recap of probability and statistics with applications. Topics include basic combinatorics, random variables, expected values and probability distributions</p> <p><i>Reading</i> Chapter</p>	Tuesday - lecture and exercise according to schedule in AIS
6	<p>Classification: Introduction to qualitative classification. Classifiers: logistic regression, naive Bayes, K-nearest neighbors, generalized linear models. Classification concepts: overfitting, confusion matrix, specificity and sensitivity of the analysis.</p> <p><i>Reading</i> Gareth et al. (2013) An Introduction to Statistical Learning: With Applications in R. Chapter 4</p>	Tuesday - lecture and exercise according to schedule in AIS
7	<p>The Nature of Econometrics and Economic Data: What is Econometrics and why do we need it? Steps in empirical economic analysis. The structure of economics data. Causality and the notion of “ceteris paribus” in econometric data. Definition of simple regression model.</p> <p><i>Reading</i> Wooldridge (2012) Introductory Econometrics: A Modern Approach Chapter 1 Chapter 2 (2.1)</p>	Monday 13.30-15.00 CET - lecture and Friday 14.15-15.30 CET - exercise
8	<p>Regression Analysis with Cross-Sectional Data: Part 1: Ordinary Least Squares. Properties of Ordinary Least Squares. Assumptions of the simple regression model.</p> <p><i>Reading</i> Wooldridge (2012) Introductory Econometrics: A Modern Approach Chapter 2 (2.2-2.5)</p>	Monday 13.30-15.00 CET - lecture and Friday 14.15-15.30 CET - exercise

Week	Topic Description	Day
9	Regression Analysis with Cross-Sectional Data: Part 2: Multiple regression analysis. Hypothesis testing. Reporting regression results. <i>Reading</i> Wooldridge (2012) Introductory Econometrics: A Modern Approach Chapter 3 Chapter 4 (4.2-4.6)	Monday 13.30-15.00 CET - lecture and Friday 14.15-15.30 CET - exercise
10	Violation of Assumptions: Heteroskedasticity and serial correlation. Endogeneity: Omitted variables, selection, and reverse causality. Instrumental variables. <i>Reading</i> Wooldridge (2012) Introductory Econometrics: A Modern Approach Chapter 3 (3.3) Chapter 8 Chapter 15 (15.1-15.2) <i>Paper to be discussed</i> Acemoglu, D., & Johnson, S. (2005). "Unbundling institutions." <i>Journal of Political Economy</i> , 113(5), 949-995.	Monday 13.30-15.00 CET - lecture and Friday 14.15-15.30 CET - exercise
11	RDD and Intro to Panel Data: Regression Discontinuity Design. Introducing time dimension. Fixed effects model. <i>Reading</i> Wooldridge (2012) Introductory Econometrics: A Modern Approach Chapter 14 (14.1) Angrist and Pischke (2009) Mostly Harmless Econometrics Chapter 6 (pp. 189–195) <i>Paper to be discussed</i> Carpenter, C., & Dobkin, C. (2009). "The effect of alcohol consumption on mortality: Regression discontinuity evidence from the minimum drinking age." <i>American Economic Journal: Applied Economics</i> , 1(1), 164-182.	Monday 13.30-15.00 CET - lecture and Friday 14.15-15.30 CET - exercise
12	Further Panel Data Methods: Random effects model. Differences-in-Differences. <i>Reading</i> Wooldridge (2012) Introductory Econometrics: A Modern Approach Chapter 14 (14.2) Angrist and Pischke (2009) Chapter 5 (pp. 169–182) <i>Paper to be discussed</i> Card, D., Krueger, A. B. (1994). "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." <i>The American Economic Review</i> , 84(4), 772–793.	Monday 13.30-15.00 CET - lecture and Friday 14.15-15.30 CET - exercise