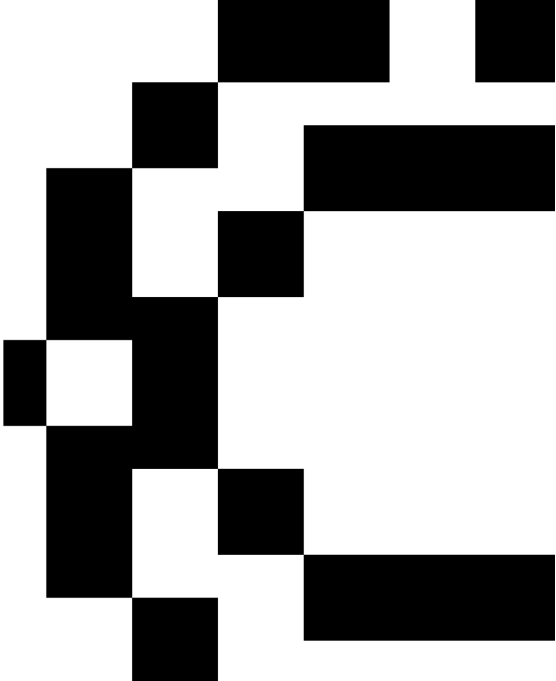
NOVA IMS INFORMATION MANAGEMENT SCHOOL

Syllabus

Statistics for Data Science

Bruno Damásio
bdamasio@novaims.unl.pt
@bmpdamasio

2025/2026



General info

- Lecturer:
  - Bruno Damásio (bdamasio@novaims.unl.pt); Office/Hours: Office 25 (2^{nd} floor) /under appointment
- Teaching assistants:
  - Beatriz Monteiro (bmonteiro@novaims.unl.pt); Office/Hours: Room 231 (2nd floor)/under appointment
  - Carolina Pinto (cpinto@novaims.unl.pt); Office/Hours: Room 231 (2nd floor)/under appointment
  - Carolina Shaul (cshaul@novaims.unl.pt); Office/Hours: Room 231 (2nd floor)/under appointment
  - Carolina Vasconcelos (cvasconcelos@novaims.unl.pt); Office/Hours: Colégio Almada Negreiros - Office 136/under appointment

Goals

The objective of data analysis is to extract the relevant information contained in the data which can then be used to solve a given problem. The purpose of this course to provide and solid background in exploratory data analysis, statistical inference, econometrics, and causal inference namely:

- Understand the basic tools to describe data (in **R** software) and the most common statistical distributions;
- Understand the main concepts of statistical inference and learn how to test a statistical hypothesis;
- Formulate and specify econometric models to interpret "economic" phenomena for sectional data and panel data;
- Recognize and understand completely the main econometric models used in cross-sectional and panel data;
- Understand the basic tools to perform causal inference.

Topics

1. Basic concepts in statistics;
2. Multiple linear regression model;
3. Panel data methods;
4. Causal inference methods.

1. Basic concepts in statistics

1. Describing and exploring data;
2. Statistical Distributions;
3. Point estimation and hypothesis testing.

2. Multiple linear regression model

1. Estimation (OLS), functional form and interpretation;
2. Inference on the parameters;
3. Testing for heteroskedasticity and functional form misspecification.

3. Panel data methods

1. Pooled OLS;
2. Fixed and Random effects estimator;
3. Hausman test.

4. Causal inference methods

1. Potential Outcomes;
2. Directed Acyclic Graphs;
3. Difference-in-Differences.

Assessment and grading

- The final grade depends on the project grade (PG) and the final exam grade (EG) (1st or 2nd rounds).
- The final grade (FG) is given by the formula

$$FG = \max\{0.5PG + 0.5EG; EG\}$$

- The formula implies the following:
  - If the final exam grade is higher than the project grade, the final grade is given exclusively by the final exam grade. Otherwise, the final grade is given by the formula $0.5PG + 0.5EG$.
  - You may decide not to do the project and thus the final grade depends solely on the final exam.
  - For approval in the course it is a necessary condition but not sufficient to obtain a minimum final grade of 9.5. It is also essential to have a minimum classification of 8 values at all components of evaluation.

Assessment and grading: Remarks on the Exam

1. Attendance to at least one of the exams is mandatory.
2. In the exams, the students may only consult a formula sheet (2 A4 manuscript pages); and a non graphical and without alphanumeric keyboard calculator.
3. Exam is not about **R** at all. Questions will approach the set of techniques lectured.
4. Of course, **R** outputs might be presented for interpretation of results.

Assessment and grading: Remarks on the Project

1. You are asked choose a data set and apply some of the models introduced in the course.
2. Typically, when students start doing research and want to present it in conferences, they do posters. The deliverable you have to submit and present must be a POSTER.
3. You may find plenty of resources on the internet where you can learn about the "practices" of communicating ideas and results on posters.
4. After submitting their projects, students may be randomly invited to discuss their work individually. This discussion may change the final grade.
5. The written assignment can be done in groups composed in the maximum of 4 elements.

Assessment and grading: Remarks on the Project

The projects must be submitted through moodle until:

The day of the first season exam at 23h59

Important: Projects submitted after the deadline will be accepted. In these circumstances, the final classification (y) results from the application of a penalty to the initial grade (x) depending on the number of days late (d) according to the following formula:

$$y = xe^{-\frac{d}{12}}$$

Main bibliography

- Lock, Robin, H., Lock, P. F., Morgan, K.L., Lock, E.F.; Lock, D.F. (2017) Statistics: unlocking the power of data. Second edition, Wiley.
- Wooldridge, J. M., Introductory econometrics: A modern approach, 6th Edition. South-Western, Cengage Learning, 2016;
- Cunningham, S., Causal Inference: The Mixtape, Yale University Press, 2021.
- Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., Vermeersch, C. M. J., 2016. Impact Evaluation in Practice, Second Edition. Washington, DC: World Bank.

Complementary bibliography i

- Heiss, F., Using **R** for Introductory Econometrics, 1st Edition; CreateSpace (Independent publishing platform), 2016.
- Stock, J. H and Watson, M. W., Introduction to Econometrics, 3rd. Edition, Pearson, Addison Wesley, 2011.
- Wooldridge, J.M., Econometric Analysis of Cross Section and Panel Data,2nd Edition, The MIT Press, 2010.
- Greene, W. H., Econometric analysis, 7th edition, Pearson , 2012;

Complementary bibliography ii

- Angrist, J. D., and Pischke, J.S., Mostly Harmless Econometrics. Princeton University Press, 2008.
- Angrist, J. D., Pischke, J. S., Mastering' metrics: The path from cause to effect. Princeton university press, 2014.
- Huntington-Klein, N., The effect: An introduction to research design and causality. Chapman and Hall/CRC, 2021.

Complementary bibliography iii

- Wickham H, Grolemund G., R for data science: import, tidy, transform, visualize, and model data. O'Reilly Media, Inc., 2016.
- Adler, J., R in a nutshell: A desktop quick reference, 1st Edition, O'Reilly Media, 2010;
- Teetor, P., R cookbook: Proven recipes for data analysis, statistics, and graphics, O'Reilly Media, 2011.