

MGT-499

Statistics and data science

Chiarotti Edoardo, Gallea Quentin, Thurm Boris

Cursus	Sem.	Type
Managmt, dur et tech	MA1	Obl.

Language of teaching	English
Credits	4
Withdrawal Session	Unauthorized Winter
Semester	Fall
Exam	During the semester
Workload	120h
Weeks	14
Hours	4 weekly
Courses	2 weekly
Exercises	2 weekly
Number of positions	50

It is not allowed to withdraw from this subject after the registration deadline.

Remark

Courses given on UNIL Campus.

Summary

This class provides a hands-on introduction to statistics and data science, with a focus on causal inference and applications to sustainability issues using Python.

Content

- Exploratory Data Analysis: Data acquisition and cleaning; Descriptive Statistics; Data Visualization; Data Ethics, Bias, and Fairness
- Causal Inference: Linear Regression; Fixed effects; Non-linear Regression; Randomized Control Trial; Regression Discontinuity Design; Difference-in-Differences; Instrumental Variables
- Applications in Python

Keywords

Data Science, Statistics, Econometrics, Causal Inference, Regression, Python

Learning Prerequisites**Recommended courses**

- Analysis
- Algebra
- Probability and statistics
- Econometrics
- Introduction to Python

Important concepts to start the course

- Basic probability and statistics knowledge (random variable, expectation, mean, conditional and joint distribution, independence, Bayes' rule, central limit theorem)
- Basic linear algebra (matrix multiplication, system of linear equations)
- Multivariate calculus (derivative w.r.t. vector and matrix variables)
- Basic programming skills (labs will use Python, basic knowledge will help)

Learning Outcomes

By the end of the course, the student must be able to:

- Describe the main pitfalls behind data analysis
- Investigate dataset, and the problems and bias behind the data
- Explore and clean datasets
- Visualize datasets
- Decide which statistical/econometrics methods to use for a given problem
- Implement these methods in Python
- Estimate model parameters from empirical observations and confidence bounds
- Test hypothesis

Transversal skills

- Plan and carry out activities in a way which makes optimal use of available time and other resources.
- Demonstrate the capacity for critical thinking
- Use a work methodology appropriate to the task.
- Access and evaluate appropriate sources of information.

Teaching methods

- Lectures
- Exercice sessions: coding lab sessions
- Group projects

Expected student activities

The students are expected to:

- attend and actively participate in lectures and lab sessions
- work on the weekly theory and coding exercises
- complete assignment (graded)
- collaborate on group projects making use of the theory learned during lectures and code developed during lab sessions (graded)

Assessment methods

- Assignments: 30% (personal)
- Group projects: 70%

Supervision

Office hours	No
Assistants	Yes
Forum	No
Others	Slack channel

Resources

Virtual desktop infrastructure (VDI)

No

Bibliography

- [not mandatory] *Mostly Harmless Econometrics*, by Angrist, Josh and Steve Pischke (2008), Princeton University Press, EPFL library
- [not mandatory] *Python Data Science Handbook: Essential Tools for Working with Data*, by Jake VanderPlas (2016), O'REILLY, EPFL library
- [not mandatory] *Introduction to Computation and Programming Using Python, Revised And Expanded Edition*, by John V. Guttag (2013), The MIT Press, MIT Press
- [not mandatory] *A Primer on Scientific Programming with Python*, by Hans Petter Langtangen (2016), Springer, Springer Link

Ressources en bibliothèque

- [Mostly Harmless Econometrics / Angrist](#)
- [Python Data Science Handbook / VanderPlas](#)
- [Introduction to Computation and Programming Using Python / Guttag](#)
- [A Primer on Scientific Programming with Python / Langtangen](#)

Notes/Handbook

Slides will be made available on a Moodle page. Notebooks will be made available in a GitHub repository.

Moodle Link

- <https://go.epfl.ch/MGT-499>

Prerequisite for

Data Science and Machine Learning (MGT-502)