University of St.Gallen

Course and Examination Fact Sheet: Spring Semester 2019

8,272: Big Data Statistics for R and Python

ECTS credits: 4

Overview examination/s

(binding regulations see below)
Decentral - Group examination paper (all given the same grades) (50%)
Decentral - Oral examination (individual) (50%, 15 mins.)

Attached courses

Timetable -- Language -- Lecturer
8,272,1.00 Big Data Statistics for R and Python -- Englisch -- Matter Ulrich , Fengler Matthias

Course information

Course prerequisites

'A Brief Introduction to Programming with R' (Master Integration Week), solid knowledge in statistics.

Course content

Short summary

This course introduces students to the concept of Big Data in the context of empirical economic research. In the first part of the course, students learn about the computational constraints underlying Big Data Analytics and how to handle them in the statistical computing environment R (local and in the cloud). Revisiting basic statistical concepts, we look at each step of dealing with large data sets in applied statistics in economic research (storage/import, transformation, visualization, aggregation). In the second part of the course, students are introduced to key methods in multivariate statistics using Python.

Description

The increasing size of datasets in empirical economic research (both in number of observations and number of variables) offers new opportunities and poses new challenges for economists. 'Big Data' is discussed as the new 'most valuable' resource in highly developed economies, driving the development of new products and services in various industries. Extracting knowledge from large data sets is increasingly seen as a strategic asset for firms, governments, and NGOs. Successfully navigating the data driven economy presupposes a certain understanding of the technologies and methods to gain insights from Big Data. This course introduces students to the basic concepts of Big Data Analytics and the basics of Multivariate Statistics to gain insights from large and complex data sets. The course combines conceptual/theoretical material with the practical application of the concepts with the open source programming languages Python and R. Thereby, students will acquire the basic skillset of analysing large data sets both locally and by means of external memory in the cloud. The practical applications of the learned techniques are focused on empirical research in economics and the social sciences. The course consists of two parts: In the first part (taught by Prof. Dr. Ulrich Matter), students learn about the computational constraints underlying Big Data Analytics and how to handle them in the statistical computing environment R. Revisiting basic statistical concepts, the course then focuses on each step of dealing with large data sets in applied statistics in economic research (storage/import, transformation, visualization, aggregation). In the second part of the course (taught by Prof. Dr. Matthias Fengler), students are introduced to key methods in multivariate statistics in order to extract insights from Big Data. This part covers clustering, dimension reduction and factor analysis. The main computing environment of the second part is Python.

Course objectives

- Students will know the concept of Big Data in the context of economic research.

- Students will understand the technical challenges of Big Data Analytics and how to practically deal with them.

- Students will know the basic statistical techniques of clustering, dimensionality reduction, and factor models.

- Students will know how to apply the relevant R packages and programming practices to effectively and efficiently handle large data sets.

- Students will know how to apply Python for Big Data Analytics.

Course structure

Part I: Big Data: Basic Concepts and Applications in R (UM)

1. Introduction: Big Data, Data Economy (Concepts). M: Walkowiak (2016): Chapter 1
2. Programming with Data, R Refresher Course (Concepts/Applied). M: Walkowiak (2016): Chapter 2
3. Computation and Memory (Concepts)
4. Cleaning and Transformation of Big Data (Applied). M: Walkowiak (2016): Chapter 3: p. 74-118.
5. Aggregation and Visualization (Applied: data tables, ggplot). M: Walkowiak (2016): Chapter 3: p. 118-127. C: Wickham et al. (2015), Schwabish (2014).
6. Distributed Systems, MapReduce/Hadoop with R (Concepts/Applied). M: Walkowiak (2016): Chapter 4
7. Data Storage, Databases Interaction with R. M: Walkowiak (2016): Chapter 5

Part II: Big Data: Multivariate Statistics in Python (MF)

1. Multivariate random variables and distributions M: Härdle, Simar (2015): Chapter 4-5
2. Clustering M: Härdle, Simar (2015): Chapter 13
3. Principal Component Analysis M: Härdle, Simar (2015): Chapter 11
4. Factor Models M: Härdle, Simar (2015): Chapter 12
5. Summary/Q&A

Course literature

Main textbooks

Walkowiak, Simon (2016): *Big Data Analytics with R.* Birmingham, UK: Packt Publishing.

Härdle, W. and L. Simar (2015): Applied Multivariate Statistical Analysis, Springer-Verlag.

Needham (2017): Python: For Beginners: A Crash Course Guide To Learn Python in 1 Week.

Journal articles and additional books

Wickham, Hadley and Dianne Cook and Heike Hofmann (2015): Visualizing statistical models: Removing the blindfold. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. 8(4):203-225.

Schwabish, Jonathan A. (2014): An Economist's Guide to Visualizing Data. *Journal of Economic Perspectives*. 28(1):209-234.

Additional course information

--

Examination information

Examination sub part/s

1. Examination sub part (1/2)

Examination time and form

Decentral - Group examination paper (all given the same grades) (50%)

Remark
group size: 2 to max 3. people.

Examination-aid rule
Term papers

- Term papers must be written without anyone else's help and in accordance with the known quotation standards, and they must contain a declaration of authorship.
- The documentation of sources (quotations, bibliography) has to be done throughout and consistently in accordance with the APA or MLA standards. The indications of the sources of information taken over verbatim or in paraphrase (quotations) must be integrated into the text in accordance with the precepts of the applicable quotation standard, while informative and bibliographical notes must be added as footnotes (recommendations and standards can be found, for example, in METZGER, C. (2017), Lern- und Arbeitsstrategien (12th ed., Cornelsen Schweiz).
- For any work written at the HSG, the indication of the page numbers both according to the MLA and the APA standard is never optional.
- Where there are no page numbers in sources, precise references must be provided in a different way: titles of chapters or sections, section numbers, acts, scenes, verses, etc.
- For papers in law, the legal standard is recommended (by way of example, cf. FORSTMOSER, P., OGOREK R. et SCHINDLER B. (2018, Juristisches Arbeiten: Eine Anleitung für Studierende (6. Auflage), Zürich: Schulthess, or the recommendations of the Law School).

Supplementary aids
.

Examination languages
Question language: English
Answer language: English

---

2. Examination sub part (2/2)

Examination time and form
Decentral - Oral examination (individual) (50%, 15 mins.)

Remark
--

Examination-aid rule
Extended Closed Book
The use of aids is limited; any additional aids permitted are exhaustively listed under "Supplementary aids". Basically, the following is applicable:

- At such examinations, all the pocket calculators of the Texas Instruments TI-30 series and mono- or bilingual dictionaries (no subject-specific dictionaries) without hand-written notes are admissible. Any other pocket calculator models and any electronic dictionaries are inadmissible.
- In addition, any type of communication, as well as any electronic devices that can be programmed and are capable of communication such as notebooks, tablets, PDAs, mobile telephones and others, are inadmissible.
- Students are themselves responsible for the procurement of examination aids.

Supplementary aids
.

Examination languages
Question language: English
Answer language: English

Examination content

- Take-home exercises (group task): Application of basic concepts in R when working with big data. Conceptual questions related to the application.
- Oral exam: Multivariate statistics (methods/concepts).

Examination relevant literature

Main textbooks

Walkowiak, Simon (2016):*Big Data Analytics with R*.Birmingham, UK: Packt Publishing.

Härdle, W. and L. Simar (2015): Applied Multivariate Statistical Analysis, Springer-Verlag.

Needham (2017): Python: For Beginners: A Crash Course Guide To Learn Python in 1 Week.

Journal articles and additional books

Wickham, Hadley and Dianne Cook and Heike Hofmann (2015): Visualizing statistical models: Removing the blindfold.*Statistical Analysis and Data Mining: The ASA Data Science Journal*. 8(4):203-225.

Schwabish, Jonathan A. (2014): An Economist's Guide to Visualizing Data.*Journal of Economic Perspectives*. 28(1):209-234.

Please note

We would like to point out to you that this fact sheet has absolute priority over other information such as StudyNet, faculty members' personal databases, information provided in lectures, etc.When will the fact sheets become binding?

- Information about courses and examination time (central/decentral and grading form): from the start of the bidding process on 24 January 2019
- Information about decentral examinations (examination-aid rule, examination content, examination relevant literature): after the 4th semester week on 18 March 2019
- Information about central examinations (examination-aid rule, examination content, examination relevant literature): from the start of the enrolment period for the examinations on 08 April 2019

Please look at the fact sheet once more after these deadlines have expired.