Module Description, available in: EN

Advanced Statistical Data Analysis

General Information

Number of ECTS Credits

3

Module code

TSM_AdvStDaAn

Valid for academic year

2024-25

Last modification

2023-09-22

Coordinator of the module

Andreas Ruckstuhl (ZHAW, rkst@zhaw.ch)

Explanations regarding the language definitions for each location:

- Instruction is given in the language defined below for each location/each time the module is held.
- Documentation is available in the languages defined below. Where documents are in several languages, the percentage distribution is shown (100% = all the documentation).
- The examination is available 100% in the languages shown for each location/each time it is held.

	Lausanne			Lugano	Zurich		
Instruction					**X** E 100%		
Documentation					**X** E 100%		
Examination					**X** E 100%		

Module Category

TSM Technical scientific module

Lessons

2 lecture periods and 1 tutorial period per week

Entry level competences

Prerequisites, previous knowledge

- Basic calculus and linear algebra
- Basic knowledge in probability, statistical inference and regression analysis on the level of Devore, Farnum and Doi, “Applied Statistics for Engineers and Scientists”, 2014 Cengage Learning.
- User knowledge of R, MATLAB, Python or any other statistical software.

Brief course description of module objectives and content

One of the most used (statistical) models for inferential data analysis is the linear regression model. But it is restricted to a Gaussian distributed response and a linear function for linking the linear combination of predictors with the expected response. Generalized Linear and Additive Models (GLM, GAM) allow us to relax some of these restrictions by specifying a more general set of response distributions and non-linear link functions. Hence we can analyse a wider variety of real world phenomenon such as counts, binary outcomes proportions and amounts (i.e. non-negative real-valued data). The aim of this modelling approach is to better understand the response outcome induced by the predictors based on the available data,

allowing for better and more informed interpretation of the phenomenon. The first part of this course will provide an overview over the GLM/GAM approach and will detail many benefits and a few pitfalls.

The second part of this course introduces to the basic concepts of causality. Many statistical and machine learning methods (including the statistical models learned in the first part of this module) are about association rather than causation. We will have a closer look at how causal effects are mathematically defined and what assumptions about data and model are necessary for drawing causal conclusions (e.g. interventions, instrumental variables, counterfactuals). In a first step, we start with the definition of causality and introducing graphical models. This enables us in a second step to estimate causal effects and interfere on causal relationships. In particular, the course introduces to structural equation models.

Aims, content, methods

Learning objectives and competencies to be acquired

- The students are able to analyse data by Generalized Linear and Additive Models (GLM and GAM) and understand the benefits that these model approaches offer for the analysis of normally and non-normally distributed response variables.
- The students understand when causal reasoning is important and what regression and machine learning are actually doing. They understand the importance of the data generating process. They can determine causal effects from observational data using graphical models.
- The students acquire a comprehensive overview how the open source statistical environment R is used and are able to perform a data analysis applying the techniques introduced in the course on real data sets.

Module content with weighting of different components

First Part (8 weeks):

- Review of the concepts of multiple linear regression analysis with respect to inference, prediction, model evaluation and model building. Introducing some advanced topics in linear regression modelling. (3 weeks)

- Extending the linear regression model to generalized linear and additive models including logistic, Poisson, and Gamma regression. Revise inference, evaluation and variable selection for such models. (5 weeks)

Second Part (6 weeks):

- Pitfalls of drawing conclusions from observational data; Simpson's paradox and its implications. Causation versus association, When is causal reasoning important? What is regression and machine learning actually doing?
- Introducing the concept of instrumental variables and interventions, determining causal effects from observational data using the most recent approaches such as causal graphical models or structural equation models, Estimating total and direct causal effects.

Both parts:

- The contents listed are illustrated with used cases from the industrial and scientific fields. The practical work is done with the open source statistical analysis environment R.

Teaching and learning methods

Classroom teaching and practical work on computer with the statistical analysis environment R/RStudio.

Literature

Slides and lecture notes will be available in addition to recommended book chapters.

Assessment

Certification requirements

Module does not use certification requirements

Basic principle for exams

As a rule, all standard final exams are conducted in written form. For resit exams, lecturers will communicate the exam format (written/oral) together with the exam schedule.

Standard final exam for a module and written resit exam

Kind of exam

Written exam

Duration of exam

120 minutes

Permissible aids

Aids permitted as specified below:

Permissible electronic aids

- Scientific pocket calculator

- R-Studio and Statistical software R on examination laptop

Other permissible aids

- open book

Special case: Resit exam as oral exam

Kind of exam

Oral exam

Duration of exam

30 minutes

Permissible aids

No aids permitted