VRIJE
UNIVERSITEIT
AMSTERDAM

DATA ANALYSIS IN R

SYLLABUS

VU Amsterdam Summer School

6 - 20 July 2024

Any general questions for the Summer School support team? Contact amsterdamsummerschool@vu.nl.

Course Details

Title	Data analysis in R
Coordinator(s)	Meike Morren
Other lecturers	Olga Ungureanu
Study credits	3 ECTS
Form(s) of tuition	On campus
Approximate contact hours	45 hours
Approximate self-study hours	25 hours

Teaching staff (in order of appearance)

Meike Morren (1978) is an Assistant Professor of Marketing at VU University Amsterdam since 2012. With a background as a methodologist, she has a deep-seated interest in analytical methods, ranging from latent variable models to large language models (LLMs).

As the Scientific Coordinator of Sustainability at the VU she plays a pivotal role in connecting researchers across faculties, motivating them to form interdisciplinary collaborations. These collaborations are essential in tackling the climate emergency through research projects that require diverse expertise.

Olga Ungureanu is interested in: New product introductions, Digital marketing, eWOM, Social media. Her methodological interests are: Quantitative methods, Choice models, Latent-class models, Time series models, Text analysis. Currently, she teaches Digital Marketing and Metrics, Advertising, Marketing Research and Multivariate Analysis.

Course description

With the increasing use of alternative programming languages like R in data analysis, now is the time to learn their ins and outs. The large number of active programmers creating R packages makes R suitable for a range of data analysis techniques, from basic hypothesis testing to generalized linear regression, and multivariate analysis such as principal component, factor analysis, or clustering. You will apply what you have learned right away in short exercises using Rmarkdown. You will be graded using an assignment in which you will learn to deal with messy data and integrate the knowledge you obtained in the exercises. The course is highly intensive as it focuses both on interpreting statistics while also learning to program in R. The focus in the exercises and assignment is the coding in R while improving skills in interpreting statistics. With the increasing use of alternative software packages like R in data analysis, now is the time to learn their ins and outs. The large number of active programmers creating R packages makes this an up-to-date program providing a huge range of statistical analyses. This course focuses upon understanding statistical models and analyzing the results whilst learning to work with R. As well as introducing the software to newcomers, it presents basic and more advanced statistics using an overarching framework of the generalized linear model.

Learning objectives

The first week is devoted to regression analysis, and learning how to use R (i.e. run analyses, visual representation and test assumptions). We start with descriptive statistics and visual representation of data, which is the first step for most statistical analyses. We then introduce the linear regression model, a widely used model with two main purposes: modeling relationships among the variables and predicting future observations.

In the second week, we will extend the linear model to the generalized linear framework, in order to analyze non-normally distributed variables. You will learn how to reduce data dimensions using principal component analysis and how to analyze multi-item scales using confirmatory factor analysis.

Upon successful completion of the course, students will:

- be able to evaluate the quality of quantitative data sources
- be able to choose the appropriate method for analysis, depending on the data source
- be able to conduct various statistical tests
- be able to analyze data using generalized linear framework
- be able to analyze multi-item scales using principal components and factor analysis
- have developed their skills in programming

Assignments and Assessment

Every day consists of short lectures with examples, and exercises in which you apply what you have learned right away. The exercises are made in class with support of the lecturer. Note that every day the solutions are posted online and the included R code can be used in the assignment. The focus in the exercises and assignment is the coding in R and how to apply and to interpret generalized linear regression models.

After class, you are supposed to work on an assignment in which you integrate what you’ve learned in the exercises during class. In this assignment, you are asked to predict prices of avocados using a dataset ranging from 2015 until 2018 in the US. You are free to select geographical regions, time periods, and variables as you think is necessary to predict the prices of the avocados. You will conduct multiple linear regression, add a trend variable, add an interaction effect and check for confounding variables, test assumptions and interpret the results. You will also compare geographical regions using the techniques learned in the course (Cluster analysis, PCA, LDA).

This assignment will be graded. The assignment has to be made individually (using Rmarkdown) and handed in via Canvas after the course as a knitted pdf or html file. If you successfully completed the assignment, you obtain a certificate of this course. You can find a grading scheme on canvas.

Additional requirements

You have to have completed an undergraduate course in statistics, an acquaintance with basic linear algebra, the fundamentals of hypothesis testing, linear regression analysis and statistical tests such as the t-test. Nonetheless, we will briefly go over these topics again to refresh the memory. Affinity with programming is an advantage in learning R. You use a computer on which R (latest version) and R desktop (latest version) is installed. We will do all the exercises in a regular lecture room on campus where you will exclusively work on your own computer.

Course Schedule

	Week 1				
	Monday	*Tuesday*	*Wednesday*	*Thursday*	*Friday*
10:00 – 10:45	*Introduction to R*	*Distributions: How to use ggplot*	*T-test* *Dummy coding boxplot*	*Ordinary least squares*	*Multiple linear regression / standardized*
11:00 – 11:45	*Read & write datafiles* *Indexing* *Assignment*	*Descriptives & frequencies / Tidyverse*	*Simple linear regression*	*Assumptions & how to adjust for them*	*Moderation & mean centering*
11:45 – 12:15	*BREAK (Q&A with teacher)*				
12:15 – 13:00	*More R objects* *Coercion*	*Loops & functions*	*Organizing regression results* *Event ASS*	*Trend variables / autocorrelation*	*Plotting marginal effects*
13:15 – 14:30	*Q&A*	*Q&A*		*Q&A*	*Q&A*

	Week 2				
	Monday	*Tuesday*	*Wednesday*	*Thursday*	*Friday*
10:00 – 10:45	*X-test*	*Classification / LDA*	*Cluster analysis*	*Principal component analysis*	*Recap*
11:00 – 11:45	*Logit regression*	*QDA*	*Cluster analysis*	*Confirmatory factor analysis*	*Recap*
11:45 – 12:15	*BREAK (Q&A with teacher)*				
12:15 – 13:00	*Log-Likelihood-ratio test*	*Plotting decision boundary*	*Plotting cluster analysis* *Event ASS*	*Mediation analysis*	*workflow / github*
13:15 – 14:30	*Q&A*	*Q&A*		*Q&A*	*Q&A*

VU

VRIJE
UNIVERSITEIT
AMSTERDAM

VU VRIJE UNIVERSITEIT AMSTERDAM