VRIJE UNIVERSITEIT AMSTERDAM

DATA ANALYSIS IN R

SYLLABUS

VU GRADUATE WINTER SCHOOL

8 JANUARY – 12 JANUARY 2024

Practical Information

The course will take place via an online streaming platform. Further details about the streaming platform will be published on the course webpage. For questions about the contents or logistics of the course, you can reach me via email at bassi0902@hotmail.com. **In this syllabus you can find the course schedule, the requirements, and a description of the winter school Data Analysis in R program.**

Course description

R is an open-source programming language that has become very important in data science because of its versatility in the field of statistics. R is usually used when the task requires special analysis of data for standalone or distributed computing. R is also perfect for data exploration. It can be used in any kind of analysis work, as it has many tools and is also very extensible. Additionally, it is a perfect fit for big data solutions. Almost every major organizations and universities use R. Google not only uses R but has also written standards for the language that are widely accepted.
This course focuses upon understanding statistical models and analysing the results whilst learning to work with R. As well as introducing the software to newcomers, it presents basic and more advanced statistics.
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: modelling relationships among the data and predicting future observations. After that, we will extend the linear model to the generalized linear framework, in order to analyse non-normally distributed variables. Each day consists of short lectures with examples and exercises in which you apply what you have learned right away.

Learning objectives

By the end of the course, you will be acquainted with various popular R packages, and you should be able to perform different statistical analyses, writing your own functions and using attractive plots to present your data.

Literature List

Freely available via VU university library:

Dalgaard, P. (2008). Introductory Statistics with R. ISBN: 978-0-387-95475-2

Course Schedule

Day 1: Introduction

We start with explaining the basics of the R environment, and Rstudio. You will learn how to work with the main data types in R: vector, factor, matrix, list and data frames. You will learn to create variables, select cases and variables, and how to use plots. Simple functions to calculate the mean and the standard deviation are introduced.

Day 2: Data & functions

You will read a data file into R, and you will learn how to compute descriptive statistics and frequencies in R. The functions discussed last day will be applied to this survey dataset. Additionally, various loop commands which allow you to run complicated tasks on the entire dataset are discussed. We introduce vectorization as an alternative to loops. Although a loop is more intuitive, vectorization is much faster. Throughout the course, we will practice these skills in writing a function for the t-test, linear regression and the log likelihood ratio test.

Day 3: Simple regression

We will discuss how the linear model is related to the t-test. You will learn how to interpret the results with one independent dummy or interval variable, and how you can test the assumptions of linear regression.

Day 4: Multiple regression

This day builds on day 3 in which we treat simple regression. Multiple regression model additionally adds the concept of 'ceteris paribus'. We will also treat confounding and interaction effects, and when and how to use mean centering.

Day 5: Logistic regression

We will introduce logistic regression as part of the generalized linear framework. We will calculate odds ratio and discuss how it is related to the chi-square test and logistic regression. Furthermore, we will discuss the log-likelihood ratio test to compare two or more models.

Assignment(s) and Grading

At the end of the course, you are supposed to complete an assignment which is graded. The focus in the exercises and assignment is the coding in R and how to apply and interpret generalized linear regression models.

The deadline for submitting the assignment is January, Friday 19th at 23:59 CET.

Requirements & evaluation

Entry requirements

You should have completed successfully a bachelor course on statistics. It definitely helps to be acquainted with the following topics: basic levels of linear algebra, fundamentals of hypothesis testing, linear regression analysis, and statistical tests such as t-test. Nonetheless, we will briefly go over these topics again to refresh the memory. Affinity with programming is a prerequisite to learn R. Several exercises are offered throughout the course, some of which have a more advanced level. You will need a computer on which R (latest version) and Rstudio desktop (latest version) are installed.

Requirements during the course

The course consists of 5 days, in which you will practice with R using exercises.

During the course, you are required to complete an assignment which builds upon the skills you have obtained in the exercises. The exercises are made in class and given as homework, with support of the lecturer. You must complete the assignment individually. The assignment should not be longer than 1500 words (R code in the appendix) and needs to be handed in via email. If you successfully completed the assignment you obtain a certificate of this course.

Course Timetable

Please note that all times are in the CET zone (Amsterdam time zone).

Schedule					
	MON 08	**TUE 09**	**WED 10**	**THU 11**	**FRI 12**
09:00 – 10:00	Introduction to R	Data generation	Simple linear regression	Multiple linear regression	X^2 (chi-squared) test
10:00 – 10:15	BREAK				
10:15 – 11:30	Data handling	Data visualization	Comparing means: T-test	Multiple linear regression	Logistic Regression
11:30 – 11:45	BREAK				
11:45 – 13:00	Reading/writing data files	Functions and loops	Theory and assumptions of regression models	Advanced Graphics	Recap