Biostatistics & Epidemiological Data Analysis using R

1
Course Overview

Stefan Konigorski

Health Intervention Analytics Group, HPI

October 28, 2021

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Welcome

- Introduction
- Experiences with programming, R, data analysis, biostatistics?
- Expectations for course?

Course overview

- Overview: https://hpi.de/biostatistics...-using-r.html
- Lectures slides, R scripts, datasets and all other relevant materials and literature on Moodle: `https://moodle2.uni-potsdam.de/course/index.php?categoryid=2128`
- Teaching team:
  - Stefan.Konigorski@hpi.de
  - Thomas.Gaertner@student.hpi.uni-potsdam.de
  - Alaa.Noor@uni-potsdam.de
  - Jennifer.DanielOnwuchekwa@student.hpi.uni-potsdam.de

Course overview

- 6 ECTS
- Language: English

- Location: zoom
- zoom link: https://uni-potsdam.zoom.us/j/62339804745 Meeting ID 623 3980 4745, passcode 77276548
- Time: Lecture on Thursday, 15.15 - 16.45, 17.00 - 18.30, Tutorial on Tuesday 17.00 - 18.30.
- All lectures will be recorded and afterwards uploaded to Media.UP, the link posted on Moodle. Please all sign the consent sheet!
- Format: lectures, exercises, group exercises

- Always have your own laptop with R (`www.r-project.org`; R version 4.1.1 or other recent version) and RStudio installation (`www.rstudio.com`; RStudio Desktop 2019.09.0+351 or other recent version) ready on your laptop.

Requirements to get ECTS points

- Work on and hand in at least 10 of 12 homework assignments
  - Small set of questions with exercises of the material covered in the class.
  - Pass/no pass, don't count to final grade.
  - No assignment after 8th and 14th class.
  - Handed out each lecture on Thursday.
  - Due the next week on Monday 23.59 o'clock.
  - Assignments available and solutions to be handed in through Moodle.

Requirements to get ECTS points

- Work on and hand in at least 10 of 12 homework assignments
  - Small set of questions with exercises of the material covered in the class.
  - Pass/no pass, don't count to final grade.
  - No assignment after 8th and 14th class.
  - Handed out each lecture on Thursday.
  - Due the next week on Monday 23.59 o'clock.
  - Assignments available and solutions to be handed in through Moodle.

- Open book take home final exam
  - Practical data analysis of real data using R, RStudio.
  - Handed out February 16, 2022 (on Moodle).
  - To be handed in by March 9, 2022 (100% final grade) through Moodle.
  - **Individual solutions, no teamwork in groups.**

Content

Block	Class	Content	Date
R, Data manipulation, Descriptives	1	Overview & Introduction to R and data analysis	2021.10.28
	2	First steps in data analysis using R	2021.11.04
	3	Second steps in data analysis using R	2021.11.11
Epidemiology & Statistics: concepts	4	Epidemiological study designs and study planning	2021.11.18
	5	Estimation	2021.11.25
	6	Hypothesis testing	2021.12.02
	7	Missing data	2021.12.09
Data analysis w/ regression models	8	Linear regression I	2021.12.16
	9	Linear regression II	2022.01.13
	10	Regression models for binary and count data	2022.01.20
	11	Analysis of variance & Linear mixed models I	2022.01.27
	12	Linear mixed models II & Meta analysis	2022.02.03
	13	Survival analysis	2022.02.10
	14	Causal inference & Data analysis challenge	2022.02.17

(see full schedule online)

Learning goals I

- Get to know the general structure and functioning of the
  - programming language R and
  - graphical interface RStudio
  - (cf. Introduction to Programming by Prof. Arnrich).
- Get to know how different real-world datasets look like.
- Get experience how to deal with some practical challenges that can arise in the data analysis.

Learning goals II

At the end of the course, the students will be able to

- understand the main concepts of basic and selected more advanced biostatistical methods and select appropriate methods for data analysis of epidemiological studies,
- import and manipulate epidemiological and biomedical datasets in R for statistical analysis,
- perform different steps of the data analysis in R considering measurement error and missing values,
- modeling binary, count, quantitative, time-to-event, clustered and longitudinal data,
- appropriately interpret the results,
- document the analyses in a reproducible manner and report the results using R Markdown.

... i.e.: answer research questions empirically!

References: overview and help for R

- No fixed textbook.
- R introduction: `http://r-tutorial.nl`
- R download and overview: `https://cran.r-project.org`
- R packages: `https://cran.r-project.org/web/packages`
- R journal: `https://journal.r-project.org`
- RStudio cheatsheets: `www.rstudio.com/resources/cheatsheets`
- Overview: `www.rstudio.com/online-learning`

Further optional references

Introduction, overview:

- R for Data Science: `http://r4ds.had.co.nz/index.html`
- Dalgaard (2008). Introductory Statistics with R.
- Crawley (2012). The R book.

Advanced R, Graphics in R, and further references:

- Advanced R: `https://adv-r.hadley.nz/`
- Matloff (2011). The art of R programming.
- ggplot2: Elegant Graphics for Data Analysis:
  `http://had.co.nz/ggplot2/`
- Dynamic Documents with R and knitr:
  `https://github.com/yihui/knitr-book`
- Practical Regression and Anova using R:
  `https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf`

Data sets: KiGGS

- Dataset of the "Kinder- und Jugendgesundheitssurvey (KiGGS)" 2003-2006 of the Robert Koch Institute: KiGGS03_06.RData
- Sample: 17,641 children (8,985 boys, 8,656 girls) from Germany, from the baseline assessment in May 2003 - May 2006.
- 1206 variables, e.g. the modules "Kinder-Umwelt-Survey" (KUS), "Psychische Gesundheit" (Bella-Studie), "Motorik-Modul" (MoMo), "Landesmodul Schleswig-Holstein", "Ernährungsmodul" (EsKiMo, 2006).

Data sets: Medical Appointment No Shows

- Dataset to predict if patients show up to a doctor's appointment or not: NoShowdata.RData
- Sample: 110,527 doctors' appointments in different clinics in the state of Espirito Santo, Brazil, within a time frame of 3 months.
- Information: 15 variables describing the clinic, appointment, sociodemographics of patients, and whether they showed up to the appointment or not.
- Link: `https://www.kaggle.com/joniarroba/noshowappointments`

Data sets: Pima Indians Diabetes Database

- Dataset to identify variables associated with diabetes: Pima_diabetes.RData
- Sample: 768 Pima Indian US women
- Data from NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases).
- Variables: 8 predictors (number of pregnancies, BMI, insulin levels, age, ...), and information whether patient has diabetes yes/no.
- Link: `https://www.kaggle.com/uciml/pima-indians-diabetes-database`
- Background reference: Smith et al. (1988). Proc Annu Symp Comput Appl Med Care, 261–265.

Questions?