University of Massachusetts Amherst

Data Management and Analysis for Social Sciences Winter 2021

Shih-chan Dai & David, Yen-Chieh Liao

Course Description

This course aims to provide students with an overview of the theoretical and practical foundations required for managing and analyzing data in R, especially those in social sciences. It starts with an introduction to foundational programming to functional programming, data types as well as basic ideas of statistics like probability. We futher will examine the different kinds of probability theory, probability distribution, and data-generating process. Once getting a solid understanding of the probability theory, we will move on to study hypothesis testing and model estimation. More importantly, this course will combine the theoretical content with a rich set of applications so that students can polish up their coding skills in R as well as apply what they learn in real-world cases.

Schedule

Lecture TBA

Lab TBA

Office Hours TBA

- **Shih-chan Dai**
- **David, Yen-Chieh Liao (University of Essex)**

Books

Programing Application

- ***R for Everyone: Advanced Analytics and Graphics***, 2nd Edition 2017, Jared P. Lander, Addison-Wesley Data and Analytics
- ***R for Data Science : Import, Tidy, Transform, Visualize, and Model Data***, 2016, Hadley Wickham and Garrett Grolemund
- ***Advanced R***, 2015, Hadley Wickham, Howard Rosenthal, CRC

- ***Hands-On Programming with R***, Garrett Grolemund, 1st Edition 2014, O'Reilly Media
- ***An Introduction to Statistical Learning with Applications in R***, Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, 7th, 2013

Statistical Analysis

- ***An R Companion to Applied Regression***, John Fox and Sanford Weisberg, 2019, SAGE
- ***The Elements of Statistical Learning***, Trevor Hastie, Robert Tibshirani, Jerome Friedman 2009, Springer
- ***Machine Learning with R, the tidyverse, and mlr***, Hefin I. Rhys, 2019, Manning
- ***ggplot2: Elegant Graphics for Data Analysis***, 2016, Hadley Wickham, Springer
- ***R in Action Data: Analysis and Graphics with R***, 2011, Robert I. Kabacoff, Manning Publication

Lecture Notes (slides will be added after each lecture)

00 - Introduction and Data Science

- Getting R (Rstudio, environment, packages and CRAN)
- Introduction to Version Control with Git and Github

Reading: CH1, CH2, CH3 of Lander)

Slides: html | pdf

01 - Programing in R

To start programming in R, this class will help you learn the foundational components of programing and data structrue in R. - Basic R (math, variables, data types, list, matrices, arrays) - Data Structures (names, values, vectors) - Subsetting

Reading: CH4 of Lander 2017) | CH11.6 of Wickham 2015

Slides: html | pdf

02 - Programing in R

Loops are used in programming to repeat a specific block of code. In this class, you will learn to how to create functions and loop, or combine them together for analysis in R programming.

- Writing R Functions
- Control Statements (loops and iteration)
- Debugging and Condition Handling

Reading: CH8, CH9 of Lander 2017) | CH9 of Hadley 2015

Slides: html | pdf

03 - Functional Programming : Basic R

The class starts by showing a motivating example, removing redundancy and duplication in code used to clean and summarise data. We will focus onthree building blocks of functional programming: ***anonymous functions***, ***closures (functions written by functions)***, and ***lists of functions***. We will learn how to easily extract, summarize, and manipulate lists and how to export the data to desired object, be it another list, a vector.

- Anonymous Functions
- List / Vector Functions (`apply()`, `lapply()`, `sapply()`, and `tapply()`)
- Manipulating Matrices and Data Frames

Reading: CH11 of Lander 2017) | CH10, CH11 of Hadley 2015

Slides: html | pdf

04 - Visialization

This class provides a comprehensive introduction on how to plot data with R's base graphics, ggplot2 as well as lattice. - Basic Graphics - ggplot2 and ggplot2 Package Ecosystem - Interactive Graphics & Maps - R server + `Tabluea` - Additionals: Rmarkdown and Basic LaTeX

Reading: CH7 of Lander 2017) | CH 22 of Wickham and Grolemund 2016 | Hadley 2016

Slides: html | pdf

05 - Manipulation and Data Management

We will covere a variety of methods for importing and exporting data in R environment.

- Importing & Exporting Data
- Tidy and Relational Data
- Data Manipulation

Reading: CH6 of Lander 2017) | CH10, CH11 of Wickham and Grolemund 2016 | CH4, CH5 of Kabacoff 2011

Slides: html | pdf

06 - Functional Programming : Tidyverse Fundamentals

This class covers fundamentals of the Tidyverse, tidy data and how to use popular functional packages such as tibble, dplyr, ggplot2, tidyr, and purrr packages for data analysis.

- Tidyverse & Functional Programming
- Pipe, Functions, Iteration and Beyond!

Reading: CH6 of Lander 2017) | CH14-CH17 of Wickham and Grolemund 2016

Slides: html | pdf

06 - Basic Statistics

- Math Functions in R
- Descriptive Statistics

Reading: CH14, CH15 of Lander 2017

Slides: html | pdf

07 - Probability Distribution

Reading: **Slides**: html | pdf

08 - Hypothesis Testing

- Hypothesis Testing for Means
- Hypothesis Testing for Proportions

Reading: CH15 of Lander 2017

Slides: html | pdf

09 - Sampling Distribution

Reading: CH15 of Lander 2017

Slides: html | pdf

10 - Correlation, Testing and ANOVA

Reading:

Slides: html | pdf

11 - Linear Regression

This class will cover fitting and interpreting linear models, evaluating model assumptions, selecting among competing models.

- Introduction and Assumptions
- Diagnostics and Interpretation
- Formatting the Estimates: `broom`,`stargazer` and `summarytools` packages

Reading: CH8, CH9, CH10, CH11 of Kabacoff 2011

Slides: html | pdf

12 - Generalized Linear Models

This class covers formulating a generalized linear model, predicting categorical outcomes, and modeling count data. - Generalized linear models and the glm() function - Model Fit and Regression Siagnostics

Reading: CH13 of of Kabacoff 2011

Slides: html | pdf

13 - Introduction to Machine Learning

This class briefly introduces what machine learning is, supervised vs. unsupervised machine learning, how the Tidyverse play a part in training machine learning classification and regression

- Introduction to Machine Learning
- Tidying and Manipulating Data with the Tidyverse

Reading: PART of of Rhys 2019

Slides: html | pdf