University of Massachusetts Amherst

Data Management and Analysis for Social Sciences

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 concepts of statistics. We further 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 programming 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 in R

- ***R for Data Science : Import, Tidy, Transform, Visualize, and Model Data***, Garrett Grolemund and Hadley Wickham, 2016
- ***Advanced R***, Hadley Wickham 2015, Howard Rosenthal, CRC

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

Data Analysis and Statistical Learning

- ***An R Companion to Applied Regression***, John Fox and Sanford Weisberg, 2019, SAGE
- ***R for Everyone: Advanced Analytics and Graphics***, 2nd Edition, Jared P. Lander, 2017, Addison-Wesley Data and Analytics
- ***R in Action Data: Analysis and Graphics with R***, Robert I. Kabacoff, 2011, Manning Publication
- ***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 Publication

- ***ggplot2: Elegant Graphics for Data Analysis***, Hadley Wickham, 2016, Springer Publication

Lecture Notes (slides will be added after each lecture)

00 - Introduction and Data Science

Data science is an discipline that allows you to turn raw data into understanding, insight, and knowledge. The goal of this course is to help you develope skillset in R that will allow you to performe an analysis.

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

Reading: 1, 2, 3 of Lander 2017 | Preface of Grolemund and Wickham 2016

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: 4 of Lander 2017 | 11.6 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: 8, 9 of Lander 2017 | 9 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: 11 of Lander 2017 | 10, 11 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 - `lattice` pacakge - ggplot2 and ggplot2 Package Ecosystem

Reading: CH7 of Lander 2017 | 22 of Grolemund and Wickham 2016 | Hadley 2016 **Slides**: html | pdf

05 - Visialization

- Interactive Graphics & Maps
- Share: LaTeXBasics, R Markdown Documents and Shiny
- R Server + `Tableau`

Reading: CH7 of Lander 2017 | 22 of Grolemund and Wickham 2016 | Hadley 2016 **Slides**: html | pdf

06 - 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: 6 of Lander 2017 | 10, 11 of Wickham and Grolemund 2016 | 4, 5 of Kabacoff 2011

Slides: html | pdf

07 - 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: 6 of Lander 2017 | 14-17 of Wickham and Grolemund 2016

Slides: html | pdf

08 - Basic Statistics

This class covers descriptive statistics, frequency and contingency tables and essential command math functions commonly used in R.

- Math Functions in R
- Descriptive Statistics

Reading: 14, 15 of Lander 2017 | 7.1, 7.2 of Kabacoff 2011

Slides: html | pdf

09 - Probability Distribution

This class cover using R to handles all the basic necessities of statistics, including drawing random numbers and calculating distribution values (the focus of this chapter), means, variances, maxmima and minima, correlation.

- Normal Distribution
- Binomial Distribution
- Poisson Distribution

Reading: 14 of Lander 2017 **Slides**: html | pdf

10 - Hypothesis Testing

- Hypothesis Testing for Means
- Hypothesis Testing for Proportions

Reading: 15 of Lander 2017

Slides: html | pdf

11 - Sampling Distribution

Reading: 15 of Lander 2017

Slides: html | pdf

12 - Correlation, Testing

- Correlation and Covariance
- T-Tests
- Examining Relationships using Visualization

Reading: 7.2-3 of Kabacoff 2011 | 3.2 of John Fox and Sanford Weisberg 2019 | 15.3 of Lander 2017

Slides: html | pdf

13 - Analysis of Variance

This class covers how to use R to model basic experimental designs, fitting and interpreting ANOVA type models, and evaluating model assumptions

- One-way ANOVA and One-way ANCOVA
- Two-way factorial ANOVA

Reading: 9 of Kabacoff 2011

Slides: html | pdf

14 - Regression : Basic 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 in R: `broom`,`stargazer` and `summarytools` packages

Reading: 8 of Kabacoff 2011 | 4, 5 of John Fox and Sanford Weisberg 2019

Slides: html | pdf

15 - Regression : 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: 13 of of Kabacoff 2011 | 6 of John Fox and Sanford Weisberg 2019

Slides: html | pdf

16 - Additional Topic: 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. Later, we plan to cover the application of the Tidyverse on Tidymodels framework, from pre-processing (`rsample` and `recipes` packages), model training (`parsnip` package), model validation (`yardstick` packages).

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

Reading: PART of of Rhys 2019

Slides: html | pdf