Course Overview

R is an open source software environment for statistical computing and graphics and R is becoming increasingly popular both in industry and academia. This course elaborates on the basic operation of R and introduces how to conduct advanced empirical analysis using R, such as textual analysis and various machine learning algorithms.

I understand that students have many chances to learn programming or R programming specifically, such as from CS department, Duke workshops, or various online resources. Then why should you take this course? What's special about this course is that we will focus on the most useful tools for economic research: I will explain when, and why you should use R instead of other programming languages such as Matlab, Stata and Python in certain scenarios; and I will elaborate with detailed examples on how R could be used to analyze various economic problems. This will help you understand which R packages to choose from in solving economic problems among the ocean of R packages that are available.

This course is specifically designed for economics master’s and Ph.D. students. There is no pre-requisite of this course, but you will benefit more from this course if you already know at least one programing language: include but not limited to C++, Matlab, Python or Java. As a caveat, the schedule of summer course is very intensive. You are expected to attend all lectures by taking this course. Each class takes 2.5 hours and runs on every Monday, Tuesday and Thursday.

Below is a short list of what will be covered in this course:

- **Basic Topics**
  - Logic in R (data structure, boolean, Loop etc.), function in R
  - R markdown
  - Data management, dplyr, tidyr
  - Data visualization, ggplot
  - Basic analysis, regression and optimization
- **Advanced Topics**
  - Text analysis and some machine learning algorithms
  - Network analysis
  - Spatial data and spatial analysis
  - How to use API, for example how to use Google Map API in R
- **Other Topics**
  - Advanced coding practice
  - Parallel computation using R
  - Version control, Git

Course Objectives

By the end of this course you will be able to:

- Understand and apply basic operation of R: file management, objects in R, loop, classes, functions, R-markup, tests and conversion.
- Conduct basic empirical analysis using R: which includes but not limited to data cleaning, running regressions, perform various hypothesis testing. Illustrate results using R packages: ggplot2, leafflet.
• Evaluate good coding practice: creating, merging, importing and exporting datasets; working with large datasets; usage of pipeline.
• Understand the logic of advanced topics and how to use them: what is version control (Git), how to perform parallel computing with R, how to use various sampling methods in R, machine learning algorithms in R.
• Conduct advanced empirical analysis using R: textual and network analysis.

Assessment

The grades will be determined by attendance (20%), two problem sets (30%) and a project (50%). I understand that you are busy doing your own research in the summer, therefore there will be only two short problem sets. Instead, we will have multiple in-class assignments. The main purpose of in-class assignments is for you to familiarize with what is taught in each class. Therefore, attending classes is very important (actually I hope this saves your time and you won’t need to spend too much time reviewing the materials out of class). You do not need to submit in-class assignments and we will go through the solutions in class together. Regarding the final project, you can choose one of the following: You could replicate a paper of your choosing; Or you could use your own research project as the final project. See more details below.

Attendance

I understand that summer is a busy time, therefore, you can skip two classes. You need to attend 6 out of 8 classes to get full credit for attendance.

Problem sets

There will be two problem sets. The problem set is designed to help you understand and try the techniques we learn in the class. The first problem set is due at 07/08 11:59pm and the second problem set is due at 07/15 11:59pm. Each problem set will be graded at a pass or fail level, which basically means if you tried all problems in each problem set, you will get full credit. The whole point of having these problem sets is for you to familiarize with the code and concept we introduce in the class.

Projects

Choose one of the following three as your final project. You could replicate a paper of your choosing or you could use your own research project as the final project.

• If you choose to replicate a paper of your choosing: you should submit a project proposal before 07/11 11:59pm, where you should explain what the main results of the paper are, what data would be used in replicating the results (if the data is not obtainable, you could simulate your own data), which set of results you plan to replicate, why R is useful and how you are going to replicate the results.

• If you choose to use your own research project as the final project: you should submit a project proposal before 07/11 11:59pm, where you should explain what the project is about, what kind of results you want to generate, how R could be useful in generating those results.

The final project is due on 07/21 11:59pm.

Suggested Readings

There is no required text for this course, but you may find following resources helpful:

A Beginner’s Guide to R by Alain Zuur, Elena N. Ieno, Erik Meesters
Introduction to statistical data analysis with R by Matthias Kohl
R for Data Science
Text Mining with R
Advanced R

Disability Statement

Students with disabilities who believe that they may need accommodations in the class are encouraged to contact the Student Disabilities Access Office at 919.668.1267 or disabilities@aas.duke.edu as soon as possible to better ensure that such accommodations are implemented in a timely fashion.

Academic Integrity

Duke University is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Citizens of this community commit to reflect upon and uphold these principles in all academic and non-academic endeavors, and to protect and promote a culture of integrity. To uphold the Duke Community Standard:

- I will not lie, cheat, or steal in my academic endeavors;
- I will conduct myself honorably in all my endeavors; and
- I will act if the Standard is compromised.

Preferred Contact

Please do not hesitate to contact the instructor via email (yanyou.chen@duke.edu) with any questions or comments. Expect a response within one business day of email delivery.

Proposed Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Topics</th>
<th>HW</th>
</tr>
</thead>
<tbody>
<tr>
<td>07/01</td>
<td>Basics in R: R packages, basic operations, file management, objects in R, classes, tests and conversion.</td>
<td>In-class Assignment</td>
</tr>
<tr>
<td>07/02</td>
<td>Introduction to R Markdown. Functions, conditional statements and loops.</td>
<td>In-class Assignment</td>
</tr>
<tr>
<td>07/04</td>
<td>No class, happy 4th of July</td>
<td></td>
</tr>
<tr>
<td>07/08</td>
<td>Data work: creating, merging, exporting and importing datasets; strings; dates.</td>
<td>Problem Set 1</td>
</tr>
<tr>
<td></td>
<td>Visualization in R: graphs, ggplot</td>
<td></td>
</tr>
<tr>
<td>07/09</td>
<td>Advanced coding practices: pipes, tidyr, tibble and dplyr. Parallel computation Git version control</td>
<td>In-class Assignment</td>
</tr>
<tr>
<td>07/11</td>
<td>Regressions: linear regression, times-series and panel data, quantile regression. Numerical optimization.</td>
<td>In-class Assignment Project Proposal Due</td>
</tr>
<tr>
<td>07/15</td>
<td>Text analysis and machine learning algorithms.</td>
<td>Problem Set 2</td>
</tr>
<tr>
<td>07/16</td>
<td>Network analysis in R.</td>
<td>In-class Assignment</td>
</tr>
<tr>
<td>07/17</td>
<td>Spatial analysis in R. How to Use API in R.</td>
<td>In-class Assignment</td>
</tr>
<tr>
<td>07/21</td>
<td>No class, final project due date</td>
<td>Final Project Due</td>
</tr>
</tbody>
</table>