Expository Data Analysis with R

(Informal Title: Empirical Analysis of Topics in Financial Literacy)

ECOG 314 / ECON-181

About this Course

This course teaches students how to conduct data analysis and develop statistical programming skills, using the statistical programming package R, in order to gain greater understanding of a number of econometric topics and financial literacy. Students will learn how to improve their literacy in financial topics – such as banks and small business financing, mortgage finance, student loans, and payday lending and other high cost credit – as well as how they will also learn how to frame and, importantly, execute an economic research project using the statistical programming package R. While the importance of developing both financial literacy and research skills is likely self-evident to most students, developing skills in statistical programming may seem, at first blush, somewhat less exciting and potentially overwhelming or intimidating. This class is designed to help the students overcome these hurdles and leave proficient in all of these topics.

The Programming Skills You Will Develop

Conducting economics research requires both an understanding of theoretical concepts and practical knowledge of empirical methods. For most students, their previous work in econometrics and financial topics focused on the theory. This class is meant to help students develop their empirical skills.

Today most empirical work in conducted using statistical programming languages, such as Stata, R, or SAS. In this course, students will learn how to use the programming language R as a means of building their empirical toolkit. R is a free, open source programming language which has become one of the leading languages in data science and statistics, including the field of economics.

This course presents the fundamentals of data analysis using R, with emphasis on economics applications, including financial literacy topics. This course is an opportunity for students to apply many of the theoretical concepts covered in earlier econometric courses in a practical way. During the course students will improve their financial literacy and general problem solving skills through lectures, homework, and projects. By the end of the semester students will be able to use their knowledge of R to clean a data set, combine data from different sources, uncover patterns and insights in data, make predictions using statistical methods and communicate your finding through the use of tables, plots, and clear and concise prose.

After completing this course, students will be able to use R as a data analysis tool to:

- Create, read, modify and store R datasets
- Use available R packages and write custom functions
- Create figures and plots
- Perform efficient dataset manipulation
- Perform and interpret multiple linear regression
- Create, manage and share reproducible project files using R markdown packages

How we Learn Programming Skills

Programming packages are often referred to as “languages”, and like learning any new language, they can be intimidating at first. At its heart, however, statistical programming involves developing a particular discipline, or logical scheme, for working with data. The specific language (or syntax) implements the logical scheme; if you understand the ideas underlying statistical programming, you can often go to any number of sources to pin down the correct syntax regardless of the specific language used.

Like all economic research, programming is also an exercise in trial and error as well as detective work. Programs rarely work on the first pass, and require us to investigate the code, make changes, re-run the program, and repeat until the code produces the expected results. This

is true for all programmers, and students should view their hands-on experience of resolving bugs as part of their initiation into the club.

It's also worth noting that all of the instructors and TAs have been in every student's shoes as beginning programmers. The course is taught in a set of digestible modules. One warning though, the modules build upon one another, so it is critical to ensure that you have mastered the prior week's content (that means completing the homework and asking for help as needed) before each course meeting. The good news, perhaps the best news, is that you have the help of a group of supportive instructors that are invested in your success.

Instructor Team

More than 30 members of the Federal Reserve Board's staff are involved in the course in one form or another. William Ampeh, a Lead Technology Analyst at the Federal Reserve Board, developed the course content with a team of other Federal Reserve staff, and will lead the class sessions. Andrew Cohen, an Assistant Director at the Federal Reserve Board and Adjunct Professor in the Howard Economics Department is responsible for content related to financial literacy and will also coordinate logistics for the course.

More important to the success of the course is the team of Federal Reserve analysts and research assistants that serve as lecturers and teaching assistants (TAs). During course meetings, a number of TAs will circulate among students to provide them with assistance on various in-class exercises. They will also hold weekly office hours on the Howard campus, and will be available to communicate via an online Piazza website. In addition, a handful of Federal Reserve economists will deliver lectures on financial literacy topics and students will also be paired with an economist to discuss their research projects with.

Course Structure

- **Lectures –** Lectures will be held Fridays from 9:15 am to 12:15 p.m. Class time will include lectures and labs with Federal Reserve Board staff. **Class will meet in the Federal Reverse Board's building at 1801 K-Street, NW. Washington, DC.**

    - Lectures will begin promptly at 9:15 a.m. It may take some time to pass through security and set up your laptops, so we recommend that students arrive **no later than 9:00 a.m.**
    - In some cases, such as Friday holidays and the final exam/presentation times, we will need to find times to schedule course meetings that work for students.
- **Office Hours –** TAs will hold office once a week in the Howard University Economics department for two hours. In the weeks leading up to major deadlines we will offer additional sessions. The first session is Tuesday 1/17/2017, in the Economics department's conference room.
- **Github Site –** All the lectures and homework will be posted to the github site.
- **Piazza Site –** We will use a piazza site as our course wiki. This is a great format for asking questions where the TAs and instructors can respond collectively to make sure your questions are answered quickly. We encourage you to use this site to ask questions throughout the course. You will be asked to create an account and contribute to the site as part of the first homework.

Course Prerequisite

All applicants must have completed a college level course in Econometrics with a grade of **B or higher**. No prior training in programming or data science is required.

Computer

A Windows or Mac laptop is required with the following minimum configuration: 4 GB RAM or higher; 320 GB hard disk; configured to allow the installation of R and RStudio software. A limited number of loaner laptops will be made available if needed, **for in-class use only**.

Software

R and selected R packages will be the primary software for this class. R is free. Substantial instruction will be provided in lecture notes and assignments, and additional instructions will also be available in the online reference materials.

Link to Download R: Comprehensive R Archive

RStudio is the recommended R integrated development environment

RStudio Download: See https://www.rstudio.com/products/rstudio/download/preview/

RStudio is easy to install and the installation does not require any instruction. However, the following links provide additional setup and navigation guidance:

http://web.cs.ucla.edu/~gulzar/rstudio/index.html

http://dss.princeton.edu/training/RStudio101.pdf

https://support.rstudio.com/hc/en-us/sections/200107586-Using-RStudio

Recommended Optional Texts and Online Reference Materials

1. An Introduction to R, by W. N. Venables, D. M. Smith and the R Core Team. URL: https://cran.r-project.org/doc/manuals/R-intro.pdf:
2. Statistical Analysis with R. "This introduction to the freely available statistical software package R is primarily intended for people already familiar with common statistical concepts. URL: http://www.statoek.wiso.uni-goettingen.de/mitarbeiter/ogi/pub/r_workshop.pdf
3. Getting Started in Data Analysis: Stata, R, SPSS, Excel: R. A self-guided tour to help you find and analyze data using Stata, R, Excel and SPSS. The goal is to provide basic learning tools for classes, research and/or professional development. URL: http://libguides.princeton.edu/dss/R
4. Gareth James, et al. 2013. An Introduction to Statistical Learning with Applications in R. Springer site to download the corrected 6th printing pdf with access to slides and 15 hours of lecture videos. URL: http://www-bcf.usc.edu/~gareth/ISL/

5.Regression analysis by Example. A Wiley series in statistics that provides a conceptually simple method for investigating relationships among variables. URL: https://aritmatika.files.wordpress.com/2010/09/regression-by-example-4th-edition-samprit-chatterjee-ali-s-hadi.pdf

6.Stack overflow: http://stackoverflow.com/questions/tagged/r.

7.R bloggers: https://www.r-bloggers.com/

Grading

Numerical class grades will be based on:

- Homework (20%)
    - Assignments will be due by midnight the Wednesday after the class they are assigned.
    - Solutions to assignments will be sent to students the following Thursday.
    - Late assignments will lose 10% of the total for each day late.
- Participation (10%)
    - Attendance – (includes being on-time to class)
    - Good class citizenship (helping other students, contributing to the piazza site)
- Midterm (15%)
    - This will be a take home exam given a third of the way through the semester
    - This will be an individual assignment where you will not be able to ask questions of the TAs.
    - There will be no office hours while the mid-term is live.
- Final project check-ins and the final project (55%).
    - More details next week

The instructor reserves the right to amend weighting.

Topics

The first hour of most of our course meetings will cover a topic relating to financial literacy. We will then use the second hour for instruction on data analysis and programming, where the examples will typically use data relating to the content covered in the first hour of class. The order in which we cover financial literacy-related topics will be determined by the schedules of our guest lecturers. The programming content will, of course, follow a much more methodical outline, described below (The instructor reserves the right to modify this schedule as needed).

1. **Introduction to Basics**
    a. Install R and RStudio; Start RStudio, explore the features, menus and windows in RStudio, and take your first steps with R.
    b. Basic operations on "*spreadsheet-like"* objects, introduction to variables, using R as a calculator to perform simple computations, introductory use of in-built R functions (e.g., mean, sum)
    c. Introduction to basic R data types including vectors, arrays, lists, matrices, data frame and factors. Explore basic operations on the basic R data types.
    d. Comment your work, save your workspace, and exit your R session.
2. **Introduction to basic programming concepts**
    a. A make-up class and TA session to ensure students have their R session properly setup, and know how to effectively download and setup the lecture materials from Github.
    b. As introduction for students who have little prior function programming experience (i.e. only STATA). This hands-on session will introduce students to programming concepts such as flow structure, variable declaration, conditional and looping constructs using "*flash-card and white-board think out loud*" examples. These concepts will then be put together in a simple and easy to follow R script file (program).
3. **Descriptive Statistics and Exploratory Data Analysis (EDA) in R**
    a. Calculating summary statistics (min, max, mean, median, quantiles).
    b. Simple graphs (*histogram(), boxplot(), densityplot(), qqnorm()*).
    c. Transform, handle missing values, rename variables, keep and drop variables, remove duplicate observations, create summarized or aggregated tabular data with rows and columns.
    d. Read external files, keep only the variables needed, display a few lines of dataset, add comments to help later users understand what is in the dataset, and save the dataset into a native format for future use.
    e. Select variables in your dataset (by subset, column name, using logic, string search, using *$* notation and by simple name).

f. Export your dataset to some other format (Excel, Text, CSV, R dataset).
g. Clean the R workspace, load and display the saved dataset.

4. **dplyr, reshape and data.table**
    a. Introduction to efficient dataset manipulation in R
    b. Introduction to basics of how to reshape data in R. Wide and long data formats.
    c. Introduction to the basics of the *DT[i, j, by]* command in data.table, an advanced version of data.frame used to speed up data large dataset manipulation tasks
5. **R Programming, Navigating the Operating System Interface, Date and Time variables in R**
    a. Sequences and simple loops (iteration), conditional execution.
    b. Writing functions, specifying function arguments and output, writing for loops, and testing variable scope.
    c. Implement functions for selected descriptive statistics.
    d. Interactions with the operating system (*getwd(), setwd(), list.files()*).
    e. Introduction to dates and times in R
6. **Producing Graphs in R using ggplot2**
    a. Introduction to traditional graphs (line charts, bar charts, histograms, dotplots).
    b. R graphics parameters and plotting style (single and multi-plots).
    c. Scatter plots with large datasets (jittering, small points and binning).
    d. Introduction to ggplot2.
7. **Simple Linear Regression in R**
    a. Model fitting (linear regression, linear regression with categorical covariates, linear regression with interactions, predicted values, residuals).
8. **Reproducible reports using R markdown**
    Hands-on sessions on how to create project reports that can be repeated by other researchers using R markdown.