Expository Data Analysis w/ R

ECOG 314 – 001 and ECON-181-001

First Course Meeting is September 2

Introduction to Data Exploration and Analysis with R

About this Course

Conducting data and econometric analysis requires not only an understanding of theoretical concepts, but practical knowledge of how to carry out empirical work. Statistical programming packages are the means by which empirical work is conducted. In this course, students will learn how to use one such language, R, as a means of building their empirical toolkit. R has become one of the leading languages in data science and statistics. The program, which is free, is the tool of choice for data science professionals in academia, research, and industry. R users include full-time number crunchers, data curators, data visualization experts, and occasional data analysts.

The course will expose you to the basics of R as applied to cleaning up messy data, breaking up large datasets into manageable pieces, uncovering patterns, deriving insights, making predictions using statistical methods, and clearly communicating statistical findings. As a result, students will expand upon and put into practice many of the concepts they covered in econometrics.

This introductory course will meet weekly on Fridays from 9:00 am to 12:00 p.m., and will include lectures, labs with help and meeting with Federal Reserve Board research assistants. **Class will meet in the Federal Reverse Board's building at 1801 K-Street, NW. Washington, DC.** (*We are offering Metro reimbursement to and from Shaw/Howard stop to Farragut North to students, which will be paid in a lump sum in December*). We will provide instructions to registered students for clearing security and entering the building.

William Ampeh, a Lead Technology Analyst at the Federal Reserve Board has developed the course content with a team of other Federal Reserve staff, and they will run the course meetings. Andrew Cohen, an Assistant Director at the Federal Reserve Board and Visiting Professor in the Economics Department will coordinate logistics for the course.

Aim of Course

This course, informally titled, "*Introduction to Data Exploration and Analysis with R*" provides a supportive, hands-on environment for students to learn R and apply and expand their existing knowledge of econometrics by conducting statistical analysis in R.

You will be introduced to concepts and techniques that will help you learn how to program in R and master the basic syntax. A range of vocabulary will be introduced to help you solve common statistical problems. You will be encouraged to continuously practice and expand on sample programs presented in class by solving weekly assignments and by participating in a group coding project. Additional help will be available both in and out of class if needed.

What You Will Learn

Starting with variables and basic operations, you will learn how to handle data structures such as vectors, matrices, data frames and lists. You will then learn to load data from a variety of formats (including SAS, Excel and text) into R, cleanup and manipulate data (including locating missing data and transforming data), and store R datasets for future use. Next you will learn to describe and examine measurement data (descriptive statistics), fit regression models, setup simulations, construct contingency tables, and implement sampling techniques.. Finally, you will learn about the graphical capabilities of R, how to create and manage your packages with Git version control, and how to publish your package on GitHub.

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

- Create, read, modify and store R datasets
- Use available R packages and write functions in R
- Create figures and plots using R
- Perform and interpret parametric one- and two-sample tests using R
- Perform and interpret multiple linear regression using R
- Perform and interpret one-way ANOVA using R
- Create, manage and publish R packages

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.

Required Text (free)

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:

Recommended Optional Texts and Online Reference Materials

1: 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

2: 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

3: 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/

4: There are many online resources for R. Here is a Twitter feed of posts by R bloggers.

URL: https://twitter.com/Rbloggers

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. People around the world use and contribute to R. Prior knowledge of R is helpful but not required. Substantial instruction will be provided in lecture notes and assignments, and additional instructions will

also be available in the online reference materials. R documentation comes with R. Many books and free online materials address R and/or R packages.

You may use either R or Revolution R Open (now Microsoft R open)

Link to Download R: Comprehensive R Archive

Link to Download Microsoft R Open: Revolution R Open now Microsoft R Open

RStudio is the recommended R integrated development environment

RStudio Download: See https://www.rstudio.com/home/

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

Course Work

Assignments (approx. weekly); Midterm project (take home); Final project (take home).

Grading

Numerical class grades will be based on the homework (30%), midterm project (30%) and the final project (40%). The instructor reserves the right to amend weighting.

Midterm Project

This will be an individual (solo) programming project to be presented in class. You will be given two weeks to do this project. You will submit your presentation slides, a write-up containing both your R code and its results, and an explanation of how you approached the problem and why you chose that approach.

End of Semester Project

Statistical data analysis of your choice using data from UCI data repository
https://archive.ics.uci.edu/ml/datasets.html

Course Syllabus

Class Notes/Assignment

1: Introduction to Basics

Part 1: Install R and RStudio; Start RStudio, explore the features, menus and windows in RStudio, and take your first steps with R.

Part 2: Introduce the basic R data types including vectors, arrays, lists, matrices, data frame and factors. Explore basic operations on the basic R data types.

Part 3: Load the *mosaic* package, and display the functions in the *mosaic* package (using Google search or *ls("package:mosaic"*). Use *help (?)* examine the *summary* function.

Part 4: Comment your work, save your workspace, and exit your R session.

Homework 1 assigned.

2: Data Input, Management and Output

Part 1: 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.

Part 2: Clean the R workspace, load and display the saved dataset. Use *tally()* and *favstat()* from the *mosaic* package to display the distribution and relevant information on the dataset.

Part 3: Select variables in your dataset (by subset, column name, using logic, string search, using *$* notation and by simple name).

Part 4: Transform variables using the *DPlyr* package, handle missing values, rename variables, keep and drop variables, remove duplicate observations, create summarized or aggregated datasets.

Part 5: Export your dataset to some other format (SAS, MATLAB, CSV).

Homework 2 assigned.

3: R Programming and Operating System Interface

Part 1: Sequences and simple loops (iteration), conditional execution.

Part 2: Writing functions, specifying function arguments and output, writing for loops, and testing variable scope.

Part 3: Implement functions for selected descriptive statistics.

Part 4: Interactions with the operating system (*getwd()*, *setwd()*, *list.files()*).

Homework 3 assigned.

4: Generating Data

Part 1: Generate numeric sequences, factors, repetitious patterns, text to create filenames, and simple loops (iteration).

Part 2: Generate random data and simulate data that satisfy specific constraints.

Part 3: Sample data and compute statistics.

Part 4: Large dataset considerations.

Part 5: Manage R dataset, files and workspace.

Homework 4 assigned.

5: Descriptive Statistics and Exploratory Data Analysis (EDA) in R

Part 1: Calculating summary statistics (min, max, mean, median, quantiles, skewness). Centering, normalizing and scaling data.

Part 2: EDA graphs (*histogram()*, *boxplot()*, *densityplot()*, *qqnorm()*).

Part 3: Test for continuous variables (test for normality, student's *t* test, equal variances test, non-parametric tests), P-values, confidence interval estimation, power of a test.

Homework 5 assigned.

Project 1: *** * * Midterm Project Presentation * * ***

6: Simple Linear Regression and ANOVA in R

Part 1: Model fitting (linear regression, linear regression with categorical covariates, linear regression with interactions, predicted values, residuals).

Part 2: Conduct and interpret analysis of variance (ANOVA).

Homework 6 assigned.

7: Random Samples in R

Part 1: Generating data samples of a specified size in R.

Part 2: Random sampling a dataset in R using the *sample()* function.

Part 3: Speeding up processing by replacing *for loops* with matrices.

Homework 7 assigned.

8: Date and Time Variables in R

Part 1: Create and access Date-Time variables.

Part 2: Date and time operations.

Part 3: Quick introduction to time series data management in R.

Part 4: Time series plotting.

Homework 8 assigned.

9: Producing Graphs in R

Part 1: Traditional graphs (line charts, bar charts, histograms, dotplots).

Part 2: R graphics packages (ggplot, lattice) and graphics devices.

Part 3: R graphics parameters and plotting style (single and multi-plots).

Part 4: Scatter plots with large datasets (jittering, small pints and binning).

Part 5: Introduction to ggplot.

Homework 9 assigned.

10: R Package, Managing R Programs with Git and GitHub

Part 1: Build your first R package.

Part 2: Git(Hub) and (R) Markdown crash course.

Homework 9 assigned.

Project 2: * * * Final Class Project Presentation* * *