Expository Data Analysis w/ R

ECOG 314 – 001 and ECON-181-001

First Class Meets January 13

(Revised: 1:00 PM December 29, 2016)

Introduction to Data Exploration and Analysis with R

About this Course

Conducting data and econometric analysis requires both an understanding of theoretical concepts, and practical knowledge of how to conduct empirical work. Statistical programming languages are how 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 and is currently considered "a hot" programming language to learn. R is a free software environment for data science professionals in academia, research, and industry. R users include full-time number crunchers, data curators, data visualization experts, and data analysts.

The course presents fundamental computational techniques using R, with emphasis on financial literacy, economic data analysis, and general problem solving from concept, through theory and finally into realization in a computational environment using R. Specifically, this course will expose students to the basics of R as applied to cleaning data, breaking up large datasets into manageable pieces, uncovering patterns, deriving insights, making predictions using statistical methods, and clearly communicating statistical findings. Thus, 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 9:15 am to 12:15 p.m. Class time will be 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.** (*Metro reimbursement to and from Shaw/Howard stop to Farragut North to students, will be paid in a lump sum in May*).

Registered students will be provided with instructions for clearing security and entering the building.

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 Visiting Professor in the Economics Department will coordinate logistics for the course.

Course Goals

"*Introduction to Data Exploration and Analysis with R"* provides a supportive hands-on environment for students to learn R while building upon their existing knowledge of econometrics and statistics.

Students will be introduced to concepts and techniques to apply toward programming in R as they master the basic R syntax. A range of R vocabulary will be introduced to help students solve common statistical problems. Students will practice as they expand on sample programs presented in class by solving weekly assignments and participate in individual coding projects. Additional support will be available both in and out of class as needed.

Learning Objectives

Starting with variables, basic operations and computations on "*spreadsheet-like*" objects (columns denote single observations or vectors and rows denote values across multiple observations), we will explore how computations in R use the same notation and computational concepts as Excel. Expanding on the concept of ***cell referencing*** and ***cell names*** in spreadsheet computations, we will examine the five main data types used for data analysis in R (homogenous data types: atomic vectors, matrices, arrays; heterogeneous data types: lists, factors – vectors that can contain only predefined values, data frames). We will then introduce and master R's subsetting operators ([, [[, and $). Several useful applications of the subsetting operations (i.e.,

how they can be used to modify parts of an R object, and select rows and columns based on a condition) will be explored.

The course will include hands-on instructions on how to install R and RStudio integrated development environment (IDE), navigate RStudio IDE, setup and manage directories for assignment and projects, introduction to R data types, a walk through the nuts and bolts of using R, loading data from a variety of formats (e.g., SAS, Excel, plain text) into R, clean and manipulate data (e.g., locate missing data, transform data), basic tips for writing good R programs, and save the resulting R datasets for future use.

We will then describe and examine measurement data (descriptive statistics). This will be followed by an introductory lecture on a select number of efficient data storage and manipulation tools including *data.table*, *dplyr*, *reshape* and *R-SQLite* for running SQL statements on R data frames. These tools provide efficient and faster techniques for transforming data (splitting, applying, and combining), handling missing values, renaming variables, keeping and dropping variables, removing duplicate observations, and creating summarized or aggregated tabular data in R. Additionally, these tools come in handy when there is the need to summarize data by panels or collapse high-dimensional data to simpler summary statistics.

R is a "functional programming language" which provides the user with tools to write their own functions, and expand on the functionalities of functions written by other users. To examine the flexibility and reusable solutions provided by R, we will explore the techniques of writing "good" R programs and functions by modifying sample programs in-class and solving weekly programming assignments.

To prepare students for the mid-term and final projects, we will cover an introduction to manipulate "*Dates and Time*" objects in R. This will then be followed by how to produce graphs in using *ggplot2*, and an in-depth coverage on how to fit regression models in R.

Finally, we will create reproducible project files using R markdown, manage R project files with Git version control, and share project files on GitHub.

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
- Perform and interpret one-way ANOVA
- Create, manage and share reproducible project files using R markdown 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

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

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 (15%), participation (5%), midterm project (40%) and the final project (40%).

Participation include attendance, contribution to class discussions and TA sessions. 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 any of the financial and economic data repositories. This include: https://fred.stlouisfed.org/ and https://www.quandl.com/. You may also work with data provided by an economist at the Federal Reserve Board.

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: 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)

Part 3: Introduction to basic R data types including vectors, arrays, lists, matrices, data frame and factors. Explore basic operations on the basic R data types.

Part 4: Load a basic R package (i.e., *mosaic* package), and display the functions in the *mosaic* package (using Google search or *ls("package:mosaic"*). Use *help (?)* examine the *summary* function.

Part 5: Comment your work, save your workspace, and exit your R session.

Homework 1 assigned.

2: Introduction to basic programming concepts

Part 1: 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 Githup.

Part 2: As an introduction to these concepts for students who have no prior functional programming experience, 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).

Homework 2 assigned.

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

Part 1: Calculating summary statistics (min, max, mean, median, quantiles).

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

Part 3: Data Input, Management and Output

Part 4: 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 5: Select variables in your dataset (by subset, column name, using logic, string search, using *$* notation and by simple name).

Part 6: Export your dataset to some other format (Excel, Text, CSV, R dataset).

Part 7: Clean the R workspace, load and display the saved dataset.

Homework 3 assigned.

4: **dplyr, reshape and data.table**

Transform, handle missing values, rename variables, keep and drop variables, remove duplicate observations, create summarized or aggregated tabular data with rows and columns.

Part 1: Introduction to efficient dataset manipulation in R

Part 2: Introduction to basics of how to reshape data in R. Wide and long data formats.

Part 3: 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

Homework 4 assigned.

5: R Programming, Navigating the Operating System Interface, Date and Time variables in R

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()*).

Part 5: Introduction to ***Dates and Times*** in R

Homework 5 assigned.

6: Producing Graphs in R using ggplot2

Part 1: Introduction to traditional graphs (line charts, bar charts, histograms, dotplots).

Part 2: R graphics parameters and plotting style (single and multi-plots).

Part 3: Scatter plots with large datasets (jittering, small points and binning).

Part 4: Introduction to ggplot2.

Part 5: Other R graphics functions (qplot, lattice) and graphics devices.

Homework 6 assigned.

7: 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 7 assigned.

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

8: Overview of topics covered

Part 1: Summary of R commands used in earlier sessions.

Part 2: Overflow topics and efficient data manipulation in R

Part 3: Summary of descriptive data analytics in R

Part 4: Summary of linear regression models

Part 5: Panel data, random and fixed effects model.

Part 6: Overview of commonly used R plotting commands and options.

Homework 8 assigned.

9: Reproducible reports using R markdown

Hands-on sessions on how to create project reports that can be repeated by other researchers using R markdown.

Homework 9 assigned.

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