ECON 370: Economic Applications of Data Science

Alex Marsh

Fall 2023

E-mail: alex.marsh@unc.edu
Web: alexmarsh.io/teaching
Class Hours: MonWed 1:25 PM - 2:40 PM
Office Hours: MonWed 9:00 AM - 10:00 AM
Class Room: Gardner Hall 307
Office: Gardner Hall 416

THIS SYLLABUS IS A WORK IN PROGRESS AND IS INCOMPLETE!

Course Description

ECON 370 is intended to provide a broad-based introduction to numerical and data-science methods commonly used in economics. The course will first introduce students to the R programming language, assuming no prior experience. Subsequent lectures, using R, will provide students an opportunity to apply this knowledge on real-world data to achieve an economic objective. The methods used in these applications will include (but are not limited to): collecting, cleaning, merging, processing, and visualizing data, descriptive analysis, optimization, and supervised/unsupervised statistical learning. In addition, the course has an experiential component that connects students with industry leaders in economic applications of data-science through a series of on-campus events.

Course Goals

My teaching goals for this course are as follows:

1. Teach students how to competently program in R from scratch with good style and how to get help when doing so,
2. Teach students how to think and approach problems from a computational perspective,
3. Teach students basic data science skills including data visualization and basic models,
4. Imbue students with a desire to learn more about econometrics and data science.

Learning Objectives

Upon successfully completing ECON 370, students should be able to do the following:

1. Be able to write functioning, readable, and aesthetically pleasing code in the R programming language along with knowing how to get help when developing this code,
2. Given raw data, be able to manipulate the data into the correct format needed for an analysis,
3. Given data and a research question, be able to create a exploratory data visualization in the right format to get at answering the question,
4. Be able to communicate results to a non-technical audience.

Prerequisites and Requirements

ECON 101 and a declared economics major.

Course Materials

- **Textbooks:** Due to the nature of this course, there is no one textbook that will be used for all material. As such, I will be pulling readings from various free textbook online. The abbreviations in the parentheses can be used to match the reading schedule to the textbooks.

  Required Textbooks

  The following textbooks are "required" (though again, they are all available for free either online or through the UNC library):

  - ***The R Book* by Michael Crawley, Second Edition (MC)**
    - Available for free through the UNC library, just enter your Onyen information after clicking the link above. If you would like a physical copy, used copies are generally affordable and the first edition should be fine, just make sure the readings match up.
  - ***R for Data Science* by Hadley Wickham and Garrett Grolemund (WG)**
  - ***Hands-On Programming with R* by Garrett Grolemun (GG)**
  - ***An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani (JWHT)**

  Additional Textbooks

  Again, not every textbook covers the material exactly how I'd like. As well, there will be students in this class with various backgrounds. Below are some other books I recommend as additional resources or as more advanced texts:

  - *Learning R* by Richard Cotton (RC)

    - This used to be the main book I used to teach this course. However, it is not legally available online for free, and I did not entirely love how it taught R. See this as a supplement for (MC) if you would like an additional text to learn the basics of R.
  - ***Advanced R* by Hadley Wickham (HW)**
    - For those who already have some experience with programming and/or R and would like a resource for really mastering R and all it can do.

- **Tutorial Links:** There are many online resources and tutorials for learning R and R packages beyond traditional textbooks. Below is a list of links to various R package tutorials that we will be using throughout this course:
  - **dplyr tutorial**
  - **tidyr tutorial**
  - **data.table vignette**
  - **ggplot2 tutorial**
  - **R Markdown tutorial**
- **R and RStudio:** We will be using R as our primary programming language of choice for this class. R is a great language to learn on and has an excellent Integrated Development Environment (IDE) called RStudio. To install R, go to **https://cran.r-project.org** and download the correct distribution for your machine. After installing R, you can install Rstudio by going to **https://www.rstudio.com/products/rstudio/download/** and downloading the free version of RStudio Desktop.
- **Recorded YouTube Recitation Sessions:** I wish this class could have a recitation section; there is a lot of material that needs to be covered that is not suited for a traditional lecture. However, this course is not big enough (nor should it be big enough) for the Graduate School to allow us a full TA. As such, I have recorded a bunch of short lectures of the type of material we would cover in a recitation section that you are responsible for watching. These are apart of the required readings, and you should watch them. These videos will get down in the weeds and cover details that I don't want to waste class time on. To find the videos, please follow **this link.**
- **Canvas:** All announcements, materials, assignments, grades, etc will be posted on the course's Canvas site. Please visit **uncch.instructure.com** and login with your Onyen. However, I will also be posting all materials on the **Github repository** for the course. You may access them however you wish, but assignments will need to be submitted on Canvas. ***Submitted assignments will only be accepted when submitted on Canvas. No exceptions.*** Any assignment that we have will be able to be submitted on Canvas. If you are having trouble submitting an assignment, it is likely something is incorrect with the submission (e.g. unnecessarily large file, a weird file type, etc). See the course policies below for more details.

Course Content

Course Overview

The course will cover the following topics using illustrative economic applications. While this is the broad outline of the course, the exact details are subject to change based on interest and background.

- Module 1: Introduction (week 1)
  - Introduction to course, the syllabus, and R
- Module 2: R programming (weeks 1-9)
  - Data types and structures
  - Objects and the environment
  - Logic, loops, and control flow
  - Functions and miscellaneous topics (dates, regular expressions, etc)
  - Programming applications (optimization, simulation, numerical methods)
- Module 3: Data Acumen (weeks 10-12)
  - Structured and unstructured data
  - Loading, cleaning, validation, merging, and processing data
- Module 4: Data Science and Visualization (weeks 13-15)
  - Descriptive analysis
  - Visualizing and plotting data (spatial, etc)
  - Clustering, classification, etc

Module 1 is a simple introduction to the course, including the R language, what it can be used for, and how to use R and RStudio to write R scripts and develop R programs. While this will seem simple and a high-level overview, understanding the topics covered in this part will be important for doing well in the course. Students easily confuse R, RStudio, R scripts, and R Markdown documents (which we will touch on later) to the detriment of their learning, so really understanding this material is crucial for really succeeding in this class.

Module 2 is where we will spend a majority of our time. We will cover the nitty-gritty of R and how to write R scripts and develop R programs to solve problems. We will spend most of our time in base R (i.e. not using packages like tidyverse and data.table, with a few exceptions). While this is a data science course, I believe the best thing students can gain from this course is high enough competency in R so that they can confidently use it for other data science and econometrics courses in the future as well as for their personal projects, internships, and jobs after Carolina. As such, we won't spend too much time learning the interesting and complex data science models and econometric methods, but rather learning the foundational R and programming knowledge to be able to implement those models and methods later on.

Module 3 is when we will diverge from base R and start learning some of the packages tailored for working with real data such as tidyverse and data.table. I will teach both paradigms and encourage students to pick their favorite and learn it well. Both packages are great and have their strengths and weaknesses. Once competency in one has been established, it is encouraged to use both in tandem (especially since packages like dtplyr exist). What I want students to learn is how to use one of the packages and write clean, readable code with it.

Module 4 is our last module and how much time we spend here will depend upon how the pace of Modules 1-3 needs to be adjusted. Ideally, I would get to teach a the lot of fun data science methods; however, it would require too much prior knowledge of both R programming as well as math, statistics, and computer science to get to this module with enough time left to do anything really cool. That said, we will cover how to do basic descriptives with your data along with how to plot and visualize your data. Having an excellent understanding of your data and being able to describe the variation in it is very important before you implement a data science model or econometric method. The better your understanding of your data, the better you'll be able to model it or know how to estimate values from it. With whatever time is leftover, we will cover same basic unsupervised and supervised learning models.

Course Schedule

The *tentative* schedule for the course including when topics will be covered and important dates for assignments is in Table 1 at the end of this syllabus. **As mentioned, this is a tentative schedule, and it is highly likely to change as I adjust the pace of the course for y'all.** Each time a class like this is taught, it must be adjusted for the composition and backgrounds of the students. *As I made adjustments, I will keep the schedule up to date on the syllabus and on Canvas along with notifying everyone of the changes.*

Reading Schedule

The readings and YouTube recitation sections associated with each lecture are included in Table 2 at the end of the syllabus. Table 1 and 2 can be connected (i.e. merged/joined... more on this later in the course ☺) via the lecture number as the "Lecture" column is common to both tables. The remaining columns are described below:

- Required Readings: These readings are required and should be completed *before* the lecture.
- Further Readings: These readings are encouraged if you would like a different or additional reference to better understand the material covered in the required readings. Please note that the Richard Cotton (RC) book is not legally available for free.
- Advanced Readings: These are readings for those who already have experience with programming in general or the R language more specifically and would like to truly master the material we are covering. The advanced readings in the data science portion of the class at the end are for those who would like to learn more, but we simply don't have time to cover this material.

Note that the abbreviations in parentheses can be matched to the associated abbreviation at the

end of each textbook in the Course Materials section above. As well, unless otherwise stated[1], the numbers for each reading is either the entire chapter if the number is an integer or the associated sections within a chapter if the number is a decimal. For example, (MC) 1 means Chapter 1 of *The R Book* by Michael Crawley whereas (MC) 2.1-2.9, 2.11 means sections 1 through 9 and section 11 of Chapter 2 of *The R Book* by Michael Crawley.

Assessments

The following items will contribute to your overall grade:

- Problem Sets (60%): A total of five homework assignments counting equally. Assignments are due before class on the due date. These problem sets will likely be time intensive, so please plan for enough time to complete the assignment. I will make sure you have enough time between release date and submission date. There will be a 15 minute grace period in case there are any submission issues. The lowest assignment will be dropped. Late assignments will receive a score of zero.
- Participation (10%): Regular participation in class discussions and attendance at speaker events. I will be tracking participation at the start of each class. If you attend 75% of the lectures, you are guaranteed at least half (5% total) of the participation points. The rest are determined via a linear scale between 75% and 100% plus or minus how well I believe you are participating in class. This is a small class in which I can get to know all of you, and I want that to be the case. Asking questions, coming to office hours, and engaging with the material will serve you greatly both in this class, and for your time after this course. **MORE ON THIS LATER MAYBE.**
  - Attendance of Industry Talks: Part of the DATA credential involves attending industry talks. Most of these talks will be offered during our class time, so conflicts will not be an issue. However, to get the credential, your attendance is expected, and attendance to these talks is *separate* to your participation in class, even though attendance of these events will count towards your participation in this class.
- Final Project (30%): In groups of 5 or less, students will given a data task involving real-world data from a published paper in a top economics journal. The task will including using R to download, reformat, clean, and analyze the data. We will meet during the final exam period for group presentations where your group will discuss what you did, what you liked and didn't like, what you found challenging and easy, and your findings from the analysis portion of the project. The presentation will be loosely graded, but most of the grading will be of your submission of the final project. *There will also be an anonymous peer-assessment for each group member to evaluate the other members of the group.* This gives me information on each group member's opinion of the effort of the others and helps prevent free-riding.

The UNC grading scale will be used. I reserve the right to curve grades if needed, but it will only ever be in the benefit of the student. The table below shows the grading scale, which corresponds to the traditional UNC undergraduate grading scale. To read the table, the percentage in each

[1] E.g. Appendix (abbreviated as Appx.) A, B, C, and D in *Hands-On Programming with R* by Garrett Grolemun

cell corresponds to the cut-off (or minimum percentage) for each letter grade minus/neutral/plus combination. These values are *inclusive*. So an A- corresponds to a percentage greater than or equal to 90% and *strictly less than* 93%. A B+ corresponds to a percentage greater than 87% and *strictly less than* 90%. And so on.

Cut-Offs For Letter Grades			
Letter Grade	(-)	( )	(+)
A	90%	93%	NA
B	80%	83%	87%
C	70%	73%	77%
D	NA	60%	65%
F	NA	0%	NA

Course Policies

Communication Channels

Feel free to contact me with any questions. I will try to respond as soon as possible. If I do not respond within twenty-four hours, feel free to send a follow-up email. Please write your email in a professional manner with a greeting, body, and closing statement. When addressing me, "Alex" is fine. I am not a professor (only a graduate student), and I am not yet a doctor.

Please make sure to notice the inference from the communication policy time frame. ***I am not obligated to respond to emails regarding assignments that are sent within twenty-four hours of the deadline.*** There are multiple reasons behind this policy. The first is for your benefit as a student. Coding can take much longer than you initially thought and you must plan your time accordingly. Unless you are already a very proficient programmer, you are unlikely to be successful if you start working on an assignment the night before it is due. The second reason behind the policy is that I am also busy and a student. I try to have a life outside of my work as much as possible. It is unlikely that I will be checking my email if I am out with friends the night before one of your assignments is due.

No Email Submission Policy

As mentioned above, I will not accept any submission of an assignment accept on Canvas. I understand that this may seem strict, but the reason for this is that if I have to keep track of assignment submissions on Canvas and in my email, it is much more likely that something will slip through the cracks. Also, we are using Canvas as it there are many features that are useful for me and the grader when grading assignments.

During Class

It is strongly encouraged that you bring a computer to every lecture and be programming along with me and the slides. The only way you learn programming is by doing. I cannot stress this enough. I've heard some mathematicians say that math is not a spectator sport. While that is

definitely true for math, it is even more true for programming. In order to learning how to program, you have to program! So please, follow along with me during lecture.

Submitting Assignments as a Group

You are allowed to work in groups of ***three*** for the homework assignments. If you do, ***please only submit one copy of the assignment and clearly state who was in your group.*** This 1) makes grading easier on our side, 2) guarantees consistency in grading across groups members, and 3) makes clear who worked with whom.

Names on Assignments

I will be giving you a unique identifier to put on your assignments. *Please do not put your real name anywhere in the file or in the file name.* This is to remove any implicit bias from names in the grading process. This is to be fair to you and also to protect my grader.

Regrade Policy

Regrade requests must be sent via email within **one week** of the assignment being returned. ***Except for strictly clerical mistakes, a regrade request subjects the entire assignment to being regraded, not just select questions.***

While it is my goal to make grading as fair and consistent as possible, grading is ultimately subjective and the way something is marked depends upon the context of the other submissions for that assignment. The more time passes after returning an assignment, the probability of me or a grader forgetting this context increases. Furthermore, while mistakes are made during grading, they are often made in both directions (i.e. marking something too harsh *and* marking something too generously). If mistakes in one direction are being considered, it is only fair that mistakes in the other direction are considered as well. This policy is to make grading fair for everyone; not just for the student receiving a grade but also for other students in the course.

Attendance Policy

As participation is a part of your grade, attendance is *required*. Not attending lecture will hurt both your grade and your understanding of the course. If you have a University Approved Absence, please let me know so that we can make plans accordingly.

Academic Integrity and Honesty

You are required to follow the UNC Honor Code as stated. If you are unfamiliar with the honor code, please see me or visit: **https://catalog.unc.edu/policies-procedures/honor-code/**. Any violations of the honor code will be reported accordingly.

Honor Code When Writing Code

Plagiarism and cheating when it comes to writing code can be tricky to determine since so much of writing code involves Googling to figure out how to do something or when one receives an unfamiliar error. Part of the point of this class is to learn how to look things up when writing

code, and I want to *encourage* this behavior. That said, you are still expected to attribute credit where it is due. When it doubt, tell me where you got any code that you did not write by putting a link or citation in the comments of your code. Seriously, I do not mind this and would rather you tell me where you got code rather than pass it off as your own. If you are not sure, you can ask me if you need to give attribution.

Preferred Name & Preferred Gender Pronouns

Professional courtesy and sensitivity are especially important with respect to individuals and topics dealing with differences of race, culture, religion, politics, sexual orientation, gender, gender variance, and nationalities. Class rosters are provided to the instructor with the student's legal name. I will gladly honor your request to address you by an alternate name or gender pronoun for any reason. In fact, I would regularly have to make these requests in undergrad[2]. Please advise me of this preference early in the semester so that I may make appropriate changes to my records.

Discrimination and Title IX

I value the perspectives of individuals from all backgrounds reflecting the diversity of our students. I broadly define diversity to include race, gender identity, national origin, ethnicity, religion, social class, age, sexual orientation, political background, and physical and learning ability. I strive to make this classroom an inclusive space for all students. Please let me know if there is anything I can do to improve, I appreciate suggestions.

Any student who is impacted by discrimination, harassment, interpersonal (relationship) violence, sexual violence, sexual exploitation, or stalking is encouraged to seek resources on campus or in the community. Please contact the Director of Title IX Compliance (Adrienne Allison - **Adrienne.allison@unc.edu**), Report and Response Coordinators in the Equal Opportunity and Compliance Office (**reportandresponse@unc.edu**), Counseling and Psychological Services (confidential), or the Gender Violence Services Coordinators (**gvsc@unc.edu**; confidential) to discuss your specific needs. Additional resources are available at safe.unc.edu.

Accommodations for Disabilities

UNC accommodates reasonable requests for students with learning disabilities, physical disabilities, mental health struggles, chronic medical conditions, temporary disability, or pregnancy complications, all of which can impair student success. See the ARS website for contact and registration information: **https://ars.unc.edu/about-ars/contact-us**.

Counseling and Psychological Services

CAPS is committed to addressing the mental health needs of the UNC community. Please do not hesitate to reach out: **https://caps.unc.edu**

[2]My legal name is Alexander but I go by Alex.

Additional Resources

- **The Learning Center:** The UNC Learning Center is a great resource both for students who are struggling in their courses and for those who want to be proactive and develop sound study practices to prevent falling behind. They offer individual consultations, peer tutoring, academic coaching, test prep programming, study skills workshops, and peer study groups. If you think you might benefit from their services, please visit them in SASB North or visit their website to set up an appointment: **http://learningcenter.unc.edu**.
- **EconAid Center:** Additional help can be obtained through the EconAid Center. More information can be found at **https://econ.unc.edu/undergraduate/econaid/**.

Syllabus Changes

The professor reserves the right to make changes to the syllabus, including project due dates and test dates. These changes will be announced as early as possible.

Important Dates

Please check the registrar's page for important dates: add/drop, breaks, course final, etc.

- Monday, August 21, 2023: First Day of Class
- Friday, August 25, 2023: Last Day for Late Registration
- Friday, September 1, 2023: Last Day to Drop Class (No Record)
- Monday, September 4, 2023: Labor Day - No Class
- Tuesday, September 5, 2023: Well-being Day - No Class
- Monday, September 25, 2023: Well-being Day - No Class
- Friday, October 13, 2023: Last Day to Drop Class (On Record)
- Thursday, October 19, 2023: Fall Break Begins - No Class
- Monday, October 23, 2023: Fall Break Ends - Classes Resume
- Wednesday, November 22, 2023: Thanksgiving Break Begins - No Class
- Monday, November 27, 2023: Thanksgiving Break Ends - Classes Resume
- Wednesday, December 6, 2023: Last Day of Class
- Saturday, December 9, 2023: Final Exam Time

Table 1: Course Schedule

Date	Lecture	Topics	Assignments
08/21/2023	1	Intro to Course and R	
08/23/2023	2	Intro to R, Pt 1: Data Types	PS1 Released
08/28/2023	3	Intro to R, Pt 1: Cont.	
08/30/2023	4	Intro to R, Pt 2: Objects & Data	
09/04/2023	No Class	Labor Day	
09/06/2023	5	Intro to R, Pt 2: Cont.	
09/11/2023	6	Intro to R, Pt 3: Functions	PS1 Due PS2 Released
09/13/2023	7	Intro to R, Pt 4: Logic & Control Flow	
09/18/2023	8	Intro to R, Pt 5. Loops	
09/25/2023	No Class	Well-being Day	
09/27/2023	9	Industry Presentation	
10/02/2023	10	Intro to R, Pt 6: Misc.	
10/04/2023	11	Application: Simulation	
10/09/2023	12	Programming Day 1	PS2 Due
10/11/2023	13	Numerical Optimization	PS3 Released
10/16/2023	14	Numerical Optimization Cont.	
10/18/2023	15	Programming Day 2	PS3 Due Final Project Released
10/23/2023	16	Intro to Data & the tidyverse	PS4 Released
10/25/2023	17	Intro to data.table	
10/30/2023	18	Merging Data	
11/01/2023	19	Cleaning Data	
11/06/2023	20	Programming Day 3	PS4 Due PS5 Released
11/08/2023	21	Industry Presentation	
11/13/2023	22	Communicating Your Data	
11/15/2023	23	Data Visualization	
11/20/2023	24	Intro to Data Science	
11/22/2023	No Class	Thanksgiving Break	
11/27/2023	25	Unsupervised Learning	
11/29/2023	26	Supervised Learning	
12/04/2023	27	Final Project Discussion	
12/09/2023	Final	Final Exam	PS5 Due Final Project Due

Table 2: Reading Schedule

Lecture	Required Readings	Further Readings	Advanced Readings	Recitations
1	(GG) Appx. A, B, C (MC) 1 (WG) 4, 6, 8	(RC) 1, 10		**Recitation 1**
2	(MC) 2.1-2.9, 2.11	(RC) 2, 4, 5, 7	(HW) 2-4	**Recitation 1**
3	(MC) 2.1-2.9, 2.11	(RC) 2, 4, 5, 7	(HW) 2-4	**Recitation 1**
4	(GG) Appx. D (MC) 2.16, 3, 4	(RC) 3, 6, 12	(HW) 7, 12-16	**Recitation 2**
5	(GG) Appx. D (MC) 2.16, 3, 4	(RC) 3, 6, 12	(HW) 7, 12-16	**Recitation 2**
6	(MC) 2.15 (WG) 19	(RC) 6	(HW) 6, 8-11	**Recitation 5**
7	(HW) 5		(HW) 5	**Recitation 4**
8	(MC) 2.10	(RC) 8, 9	(HW) 5	**Recitation 3**
9	———————-	———————-	———————-	———————-
10	(MC) 2.12-2.13 (WG) 16	(RC) 11	(RC) 16-17	**Recitation 6**
11				**Recitation 7**
12	———————-	———————-	———————-	———————-
13				**Recitation 8**
14				**Recitation 8**
15	———————-	———————-	———————-	———————-
16	(WG) 9-12 **dplyr tutorial** **tidyr tutorial**			**Recitation 9**
17	**data.table vignette**			**Recitation 9**
18	(WG) 13			**Recitation 9**
19				**Recitation 10**
20	———————-	———————-	———————-	———————-
21	———————-	———————-	———————-	———————-
22	(WG) 26-27, 29-30 **RMarkdown**			**Recitation 11**
23	(WG) 3, 28 **ggplot2 tutorial**			**Recitation 11**
24	(JWHT) 2			**Recitation 12**
25	(JWHT) 12.1, 12.4		(JWHT) 12.2-12.3	**Recitation 12**
26	(JWHT) 3-6		(JWHT) 7-10	**Recitation 13**
27				**Recitation 14**