Environmental economics and data science
(EC 607)

Winter 2019 syllabus

Grant R. McDermott
Dept. of Economics, University of Oregon

Summary

When:	Mon & Wed, 14:00–15:50
Where:	PLC 410
Web:	https://github.com/uo-ec607
Who:	Grant McDermott
	🎓 Assistant Professor of Economics
	✉ grantmcd@uoregon.edu
	⧖ Tue & Thu, 09:00–10:00 (PLC 530)

Course description

This seminar is targeted at economics PhD students and will introduce you to advanced methods in environmental economics and data science. While some material will likely overlap with your other quantitative and empirical methods courses, this is not just another econometrics course. Rather, my goal is bring you up to speed on the practical tools and techniques that I feel will most benefit your dissertation work and future research career. This includes many of the seemingly forgotten skills — like where to find interesting data sets in the "wild" and how to actually clean them — that are crucial to any successful scientific project, but are typically excluded from core econometrics and statistics classes. We will cover topics like version control and effective project management; programming; data acquisition (e.g. web-scraping), cleaning and visualization; GIS and remote sensing products; and tools for big data analysis (e.g. relational databases, cloud computation and machine learning). In short, we will cover things that I wish someone had taught me when I was starting out in graduate school. While the data sets and materials focus will predominantly link to environmental and natural resource issues (my own fields of specialisation), the tools and methods apply broadly. Students from other fields of specialisation are thus welcome to register.

Practical matters

Class rules

Please bring your laptops to class. This will be a very hands-on course, with relatively little in the way of formal theory. Instead, we'll be working through lecture notes together in class and you'll be running code on your own machines.

Software requirements

All of the software requirements for this course are open-source and/or free. Please aim to have everything installed by the start of our first lecture. I will be available for installation troubleshooting during the first week of the quarter. If you want a detailed tutorial on how to achieve a perfect working setup, I can think of no finer guide than Jenny Bryan *et al.*'s `http://happygitwithr.com/` (see esp. sections 4 – 15).

R and RStudio

We will mainly be using the statistical programming language *R* (download here). Please make sure that you install the **RStudio IDE** too (download here).

Git and GitHub Classroom

We will also make extensive use of the **Git** version control system (download here[1]). Once you have installed Git, please create an account on **GitHub** (here) and register for an education discount to get unlimited private repos (here).[2] Now is probably a good time to tell you that I am going to run the entire course through GitHub Classroom. You will receive an email invitation to the course repo with instructions in due time, but suffice it to say that this is how we'll submit assignments, provide feedback, receive grades, etc.

Other

I will describe additional software requirements when the need arises. Lecture notes will be posted ahead of time and list the *R* packages and external libraries (if relevant) required for a particular class. Toward the latter half of the course, we'll be using various tools on the Google Cloud Platform, as well as making occasional use of other programming languages like SQL, Python and JavaScript.

---

[1]Windows users, please make sure that you download Git for Windows since that will come with Git Bash, which you'll need elsewhere in the course. More information here.

[2]GitHub recently announced unlimited free private repos for everyone. However, you are limited to three collaborators per private repo, so the education discount still makes sense.

Textbook and other readings

The nearest thing to a conventional textbook for this course is probably Garrett Grolemund and Hadley Wickham's "**R for Data Science**". I have ordered some copies at the Duck Store, but the book is available in its entirety for free online. I highly recommend this book for anyone who is interested in using *R* for their research.[3] Which, let's be honest, you should be. Only dinosaurs are using Stata now. (Don't tell the other professors. Actually, who am I kidding: TELL THEM.)

Other books that I eagerly recommend and will be drawing on occasionally include "**Advanced *R***" (Hadley Wickham, again), "**Data Visualization: A practical introduction**" (Kieran Healy), and "**Geocomputation with *R***" (Robin Lovelace, Jakub Nowosad and Jannes Muenchow). These books are all freely available online too. I may also refer you to the **STAT 545 website**, which is a course initially taught at UBC by Jenny Bryan and continues to serve as an incredible knowledge resource for all things related to *R* and reproducible research. Finally, if we get enough time to take a deep dive into machine learning, then I'll be drawing from "**The Elements of Statistical Learning**" (Trevor Hastie, Robert Tibshirani, and Jerome Friedman), which is a classic and (surprise!) also available as a free PDF online.

Taking a step back, one of the goals of this course is to make you aware of the incredible array of instruction material that is freely available online. I also want to encourage you to be entrepreneurial. In that spirit, many of the lectures will follow a tutorial on someone's blog tutorial, or involve reproducing an existing study with open source tools. Each lecture will come with a set of recommended readings, which I expect you to at least look over before class.

Evaluation and grading

Grade determination

Grades will be determined as follows:

5 ×homework assignments (15% each)	75%
2 ×short presentations (5% each)	10%
1 ×peer-review	10%
In-class participation	5%

This breakdown should (hopefully) be pretty self-explanatory. Any specific requirements will be made clear as we proceed through the course. However, here are some additional details for ~~pedants~~ people who like everything written down precisely:

---

[3]For those of you who prefer Python to *R*, Jake VanderPlas's "**Python Data Science Handbook**" is another excellent option.

Homework assignments (and/or final presentation)

Homework assignments are to be completed individually. Late submissions will not be graded. There is no final exam or project for this course. However, you have the option of swapping out one of the individual homework assignments for a final (20 min) presentation of your own research. Think of this as an opportunity to develop and refine one of your PhD projects using the tools that we will cover in this course. In particular, some of you may wish to present your second-year field paper, or a dissertation chapter idea. You are allowed to do this individually or in pairs. However, please note the following caveats: 1) You need to get prior approval from me and let me know which HW assignment you are dropping. 2) These final presentations will only be graded on content relevant to this course. (Don't present a theory paper!)

Short presentations

Most lectures have one or more key readings; see the Lecture outline at the end of this document. Each of you must give a short (5-10 min) summary presentation on at least one of these key readings. I say "at least one" because — while you will need to give two short presentations in total — you also have the option to present on an (approved) software package or tool of your choice.[4] Topics will be assigned on a first-come-first-go basis... But don't be surprised if I volunteer you for something.

Peer-review

You are going to peer-review (or reproduce) a study, project or software package. The focus here is on code and analysis, rather than framing or narrative issues. How exactly I expect you to do this will become clear after the first few lectures. The gist is that you will be using GitHub and related tools. (E.g. Cloning or forking a repo, identifying bugs or missing dependencies, issuing pull requests, and so forth. Again, these terms will make more sense once we cover them in class.) An approach that worked well last year — but depends on demand for final presentations — is that students reviewed each others' field papers. You could also choose to review any open-source project or repo, including my own. You will have 5 minutes to present your main findings/contributions and will also need to share any code changes/contributions with me.

Honesty and academic integrity

Students caught cheating or plagiarizing will automatically be assigned a zero grade. Please acquaint yourself with the Student Conduct Code at http://studentlife.uoregon.edu.

---

[4]I'll provide a list of some suggested packages and tools on the course repo.

Accessibility

If you have a documented disability and anticipate needing accommodations in this course, please make arrangements with me during the first week of the term. Please request that the Accessible Education Center send me a letter verifying your disability.

Lecture outline

Data science basics

1. Introduction: Motivation, software installation, and data visualization

2. Version control with Git(Hub)

3. Learning to love the shell

4. *R* language basics

5. Data cleaning and wrangling with the "Tidyverse"

6. Webscraping: (1) Server-side and CSS

7. Webscraping: (2) Client-side and APIs

Analysis and programming

8. Regression analysis in *R*

9. Spatial analysis in *R*

10. Functions in *R*: (1) Introductory concepts

11. Functions in *R*: (2) Advanced concepts

12. Parallel programming

Scaling up: Big data and cloud computation

13. Docker

14. Virtual machines / cloud servers (Google Compute Engine)

15. High performance computing (UO Talapas cluster)

16. Databases: SQL(ite) and BigQuery

17. Spark

18. Machine learning: (1)

19. Machine learning: (2)… Or, student project presentations (demand dependent)

20. Peer-review and student project presentations (demand dependent)

FAQ

This course looks interesting! Can I use/adapt your lecture notes for a similar course that I'm teaching at XYZ?

Sure. I've benefited greatly from other people making their teaching materials publicly available (and have tried my best to acknowledge them directly in the relevant sections of this course). Say nothing of the incredible open-source software that powers everything. I'm more than happy to pay it forward. I only ask two favours. 1) Please let me know (email/Twitter) if you do use material from this course, or have found it useful in other ways. 2) A minor acknowledgment somewhere in your own syllabus or notes would be much appreciated.

The other data science courses that I've seen all have at least one whole lecture dedicated to data visualization. Where's yours?

Every lecture in this course is dedicated to data visualization! Okay, seriously, we'll cover the basics of ggplot2 in the opening lecture (and first assignment) and consistently build upon that in subsequent weeks. Much as I'm tempted to have a standalone lecture on the topic, I have to triage because of the time constraints of a 10-week course. I don't want to run out of road before we can get to some of the big data stuff towards the end of the course. Trust me, though. There will be a *lot* of data visualization in this course.

What about regular expressions? I hear those are super important too.

100% agree and, much like data visualization, I've tried to include examples throughout the course rather than in a standalone lecture. I'm confident that you will have a solid grip of the basics by the time we get to the end of the quarter.

I hear that data scientists use Bayesian methods a lot. Will you be covering those in depth?

Sadly, no. I'm a Bayes fanboy (as my research interests will attest), but again have to think about time constraints. The good news is that running Bayesian models in *R* is super easy thanks to a multitude of packages, and you will be very well positioned to jump right into these after finishing this course. We might even get to an example or two in the lecture on regression analysis. The even better news is that Jeremy Piger teaches an excellent Bayesian course here at the UO that you should attend.

Is there anything else that you aren't covering that I should know about?

The obvious thing that springs to mind is workflow automation and analysis pipelines (make files, etc.). Again, triage rules the day. We will, however, be working extensively with R Markdown documents,

which is at least a big step in the direction of self-contained analysis. And I'm more than happy to point students in the right direction if anyone wants to learn more. (Here, here, and here are great places to start.) Another thing we won't have time for is package development and maintenance, although I don't see this class as the primary audience for that. OTOH, students will be rewarded for package contributions if they choose to do so in the peer-review section of the course.

_R_ looks cool, but I'm more familiar with Python/Julia/MatLab/etc. Can I use that instead?

Short answer: No. Longer answer: Look, I like and use a lot of those languages too, but I'm not changing my lecture notes or assignment templates for you. Plus, I really do think that _R_ makes the most sense for applied economists looking to develop their data science skills. It already has all of the statistics and econometrics support, and is amazingly adaptable as a "glue" language to other programming languages and APIs. Learning multiple languages is never a bad idea in the long run, though.

I already have a BitBucket/GitLab/etc. account. Do I still have to use GitHub?

Since I'm running this course through GitHub Classroom, yes. But good for you! (Seriously… those are great platforms too and as an open-source advocate, I fully support a plurality of tools and software options.)

On that note, do you have any advice for running a course on GitHub Classroom?

I mostly followed this excellent tutorial by Jacob Fiksel.

Remind me again: What exactly does this course have to do with _environmental_ economics?

Good question. The truth is that this is really a data science tools course taught by an environmental economist. And the really truthful truth is that getting university approval for a new course — with a different name — is a bureaucratic nightmare, compared to just modifying an existing one off the shelf. Now, having said that, we _will_ be dealing with a lot of environmental datasets and topics. From energy and pollution data to fisheries to GIS and remote sensing products. These are the products and research themes that I am most familiar with and care most deeply about. The topics in this course are also genuinely representative of the tools that I use in my day-to-day research as an environmental economist. The good news that they are very easily adaptable to other fields.