Python for Social Scientists

September 16th-27th, 2019
PhD Academy - LSE

Practical details

Dates	September 16th-27th	(no class on the 25th)
Time	10:00am-1:00pm	
Place	PhD Academy training room Lionel Robbins building, 4th floor	
Instructor	Jialin Yi PhD candidate, Department of Statistics, LSE j.yi8@lse.ac.uk	

Course Description

This two-week course will provide an overview of the tools and methods required to undertake a collaborative research project in Python. It is designed for first- and second-year PhD students in the social sciences with the ambition to conduct quantitative research in Python. No knowledge of Python is required to take this course, although familiarity with elements of programming will help students get the most out of it. The class material was created for an audience familiar with Stata, MATLAB or R, looking to transition away from proprietary softwares and to be able to undertake all aspects of a research project within a single programming environment.

By the end of the two weeks, students will be able to use GitHub to manage a collaborative research project, to use most of the econometrician's standard tools in Python, and they will have gained familiarity with standard machine learning techniques.

Prerequisites

Previous knowledge of Python is not required but will greatly help. We expect the course to be fast-paced and the instructor's time will be better used to help students assimilating the class material. If you are unfamiliar with Python, we recommend to go through the three first lectures of QuantEcon: "Introduction to Python", "the Scientific Libraries", and "Advanced Python Programming".[1] No knowledge of version control, GitHub or machine learning is required.

[1] QuantEcon is a website created by economists Thomas J. Sargent and John Stachurski, teaching social scientists how to use Python for research. Extensive accompanying lecture notes are available in .pdf format (Sargent and Stachurski, 2019).

While not a required reading, "Code and Data for the Social Sciences" by Gentzkow and Shapiro (2014) is a very informative resource. This 40-page long paper describes the best practices for collaborative research projects in the social sciences. Chapters 1, 3, 6, 7 and the appendix on code style are worth reading before the class starts.

Students should be familiar with the various statistical tools whose implementation in Python will be demonstrated during the class (see the Course Outline section below).

Computation

The course will consist of a combination of demonstrations and in-class exercises, so students are advised to bring their own laptops. If you need a laptop, the PhD Academy can provide you with a MacBook laptop for the duration of the course. There are 28 MacBooks available, but these machines will also need to be shared with other users of the PhD Academy. These laptops need to be placed back in their lockers after use, their memory will be wiped then so students will need to save their work on a fash drive or on their `H:` drive.

Please install the latest version of Python (3.7) through the Anaconda distribution, ahead of the course. The PhD Academy laptops will already have Anaconda installed. Anaconda is a free and popular platform for data scientists working with Python. It comes with all the packages we will use in the class. You can download Anaconda from here.

Course Outline

➫ indicates class demonstrations or exercises

[This section will be subject to some changes, please consider it as indicative only]

Module 1: Version control for collaborative projects 3 days		
Monday 16	*Version control*	Introduction to the course Version control, Git and GitHub ➫ Using GitHub ➫ Class exercise: using Google Cloud Platform ➫ Using Jupyter Notebooks (at home)
Tuesday 17	*Cloud Computing*	
Wednesday 18	*Intro to Python*	

Module 2: Statistics and Econometrics 5 days	
Thursday 19	*Basics of data handling*
Friday 20	*OLS, GLS, IV and NLLS*
Monday 23	*Maximum Likelihood and Limited Dependent Variable Models*
Tuesday 24	*Time Series*
Tuesday 25	–No class–
Wednesday 26	*GMM*

Module 3: Elements of Machine Learning 1 day	
Thursday 27	*Introduction to the data scientist's toolkit*

Resources

In class, the instructor will use Jupyter Notebooks for demonstrations, and students will use the Spyder scientific environment for hands-on exercises. Both can be used within Anaconda. All Notebooks, datasets, lecture notes and slides will be available on the course GitHub page.

References

Gentzkow, M. and J. M. Shapiro (2014): "Code and data for the social sciences: A practitioners guide," *University of Chicago mimeo*, https://web.stanford.edu/~gentzkow/research/CodeAndData.pdf. 2

Sargent, T. J. and J. Stachurski (2019): "Lectures in quantitative economics with Python," *mimeo*, https://lectures.quantecon.org/_downloads/pdf/py/Quantitative%20Economics%20with%20Python.pdf. 1