Thoughts for Graduate Students Starting out with R

Aaron Erlich
Department of Political Science
University of Washington
aserlich@u.washington.edu

This draft: October 2, 2014*

Many students, who are experienced with Stata wonder why they should spend the time to learn R. Often, they also don't understand what R is. They think that is an undue burden placed upon them.

So, what is R? As explained in class R is much more powerful than any other way of estimating statistical models. This is because it is not a statistical software but a program with many pieces of statistical software (programs) written for it. The difference between being a programming language and a statistical software is a key and often under appreciated point, which I think it helpful for grad students to understand.

Let me try to help the process by mentioning some important ways you need to think about R that is different than other ways of estimating statistical models on your computer.

R is an object-oriented (OO) programming language.

- That is, R is not a statistical software *per se*, rather it is a programming language that is optimized to perform matrix algebra and other operations related numerical operations that statisticians and applied research scientists find important. R like any true computing language is more powerful for general statistical analysis than canned stat packages like Stata, SPSS, and SAS, but compared to other computing languages, R is tailored to make statistical analysis easy, so it is the best of both worlds. You **should think of R as a programming language** and not a statistical software. The code (software) that estimates every model run on R you could write yourself and can (and may change) if you don't like the way it is implemented. The software written for R is free because it is generally written by academics or open source enthusiasts. Some of it works better than others! So there has to be a buyer beware message. Not all software libraries are bug free.

- R is compiled at run time (like Python). For those who don't know much about computers at this point, don't worry about this too much. But it is different than C++ and other such languages and why people sometimes don't think R is a programming language. Don't be fooled! It's a programming language! That being said, most R users are applied scientists, not computer scientists. As applied scientists our aim is saving our own time, not the world's

*Feedback on improving this material is always welcome. As always, thanks to Chris Adolph

time. That's why we write code that is "good enough" and leave compiling and optimization mostly (but not entirely) to R itself. This means with R we generally get workable code to solve our problems faster than people who work in (say) C++, but our code is usually considerably slower. That's usually the right trade-off, but if you are spending weeks waiting for your code to finish running, you may need to learn more programming skills. Even if that happens to you, learning R is a good first step.

- Since R is a programming language you can perform many operations that are not available on statistical software. You can
  - scrape data from, the web
  - parse massive text files
  - write programs to push data onto the Internet

  Of course other programming languages are better for some of these tasks (and you may need to learn these later), but almost all tasks you can perform in other programming languages you can do in R in a pinch.

Q: What does object oriented mean? Object oriented means that you can store (stuff) data in an object, which has a certain type or in R (class). Different types of objects have different types of operations that can be performed on them. These operations are usually called through functions.

Q: Can an object have more than one class? In R, the answer is yes. it is often useful to check the class of your object because often you may be trying to call a function that does not operate on the class of objects you have stored.

Q: Other Questions I