Which statistical software to use?

A reflection and review upon the use of STATA, R and Python for teaching in the social sciences

Thomas de Graaff[1]*

Abstract

Keywords
Teaching—Data—Empirical research—STATA —R—Python

[1] *Department of Spatial Economics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands*
***Corresponding author**: ✉ t.de.graaff@vu.nl; thomasdegraaff.nl

Introduction: the empirical workflow

> Econometrics is much easier without the data.
>
> Marno Verbeek

The quote above does not only apply to economics and econometrics, but to all of the social sciences as well. Empirical research—that is, dealing with data in all its forms—requires a rigorous approach. Even more so, with the increasing emphasis on openness and reproducability of research. Therefore, it is strange that in academic education there is not much guidance in choosing which tools to use and in the philosophy behing choosing an efficient and reproducable workflow.[1]

This note deals with the suitability of various software packages for applying applied econometrics in specific and data science in general.[2]. Specifically, In will focus on STATA, R and Python.[3] I will not say which tool to use. Instead, I will focus on the various strengths and weaknessess of each software package combined with it specific approach. The main criteria I will consider are the package suitability for education and how well it can be integrated in an efficient workflow. The former is mostly important for doing (small) data exercises, whilst the latter is vital for larger research project, such as theses and later on perhaps research papers.

To illustrate why the importance of reproducability, note that a typical empirical workflow in the social sciences looks as follows:

Generate data Data is read from an external source (file or online database) or is simulated.

Manipulating data This is usually the most time demanding phase[4] and includes (amongst many other things) manipulating missing data, merging data and relabeling data

Analyse data This phase includes not only standard ecometrics and statistical or machine learning techniques, but as well as graphical representations as maps and figures.

Present results Finally, documents in the forms of papers, posters, theses, or presentations have to be drafted. Note that ideally one wants to do in various formats, such as in pdf for physical paper and in html for webdisplay.

Unfortunately, all these steps do not necessarily run sequentially. Supervisors, referees, colleagues, and the future you, always want to add or delete elements to or from your research. For instance, variables have to be added, models specifications have to be checked, and 3D pie charts have to be changed in something useful.

Therefore, it is vital that all these steps are both (*i*) very well documented so that the future you can easily retrace your steps, implement changes and redo the whole research if needed, and (*ii*) well connected to each other. The latter does not necessarily entail that the whole research should be done in one software environment, but instead that the outcome of one research step (e.g., generating data) can easily serve as an input for another step (e.g., data manipulation).

In the next section I will lay out the strengths and weaknessess of the three statistical software packages according to the criteria mentioned above.

[1] There are some exceptions, see, e.g., Healy (2011).

[2] There is a difference between econometrics on the one hand and statistics on the other hand. Economics students first and foremost need to able to apply applied econometric techniques, such as presented in Angrist and Pischke (2008)

[3] I will also briefly tough upon some other packages, but these three mentioned are most likely the most used in economics, except of course for the ubiquitous Excel.

[4] There is and old saying that says that 80% of your research time goes in transforming data, while 20% is only spent on analysing the data

Statistical software packages

I review the various packages according to several criteria. There are several other that will be discussed, but these I find most important for a suitable software package to be used for teaching.

Open source The most important argument to use an open source package is reproducability. Your work is simply less accessible and thus reproducable if the code can only be run with applications that costs over €1,000.

Learning curve First and foremost, students should be able to use the package for straightforward econometric research. If that is not possible after one six-week's course, the software package is not particularly suitable.

Size of the community Nobody want to be locked in with obsolete technology. A large userbase ensures a high probability that the software package will be used and maintaned in the future as well. Moreover, all sorts of indirect effects, such as user written routines, packages and documentation, come along for free with a large community.

Usefulness outside academia Often forgotten as an argument but outside acadamic life, some applications are more used than others. And with the recent emphasis on better preparation for the labor market, this argument seems to become more importnant. By the way, the application still mostly used would be the ubiquitous Excel and its related visual basic scripts.

Flexibility

STATA
R
Python

Statistics in education

Concluding remarks

References

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Healy, K. (2011). "Choosing your workflow applications". In: *The Political Methodologist* 18.2, pp. 9–18.