

TEACHING STATISTICS WITH MATLAB: HOW AND WHY?

Chibueze Ogbonnaya

Centre for Applied Statistics Courses, Institute of Child Health,
University College London, United Kingdom.

c.ogbonnaya@ucl.ac.uk

Teaching statistics with software is a practical approach for introducing students to real world problems and how to analyse data. There are different choices of software for teaching statistics, however R and SPSS have been the most common choices. This has often caused a dilemma for students and researchers who have an engineering background and may prefer to do most of their analysis within MATLAB instead of learning a new programming language they may not use often. This paper explores lesson learned from creating a new MATLAB short course and some useful tools for teaching statistics using MATLAB.

INTRODUCTION

MATLAB (The Math Works, 2020) is an object-oriented programming language that is commonly used in academia as well as industry for scientific computation, simulation, and data analysis. Although MATLAB has been commonly used by students and researchers in the field of engineering, physics, and other physical science fields for simulations, optimization, and calculus, the capabilities of MATLAB for statistical analysis are often ignored. This often implies that researchers in engineering will often want to use R (R Core Team, 2021) or SPSS (IBM Corporation, 2021) for their statistical analysis, which can be tedious because it may require moving data from MATLAB to SPSS or R while ensuring compatibility. To improve knowledge on the statistical capabilities of MATLAB, it was important to create an introductory course on MATLAB that covers the basics of statistical data analysis using the MATLAB language. The course is targeted at students with little or no prior knowledge of MATLAB and with a basic knowledge of statistics (statistical tests, p -values, and confidence intervals) and covers loading data, MATLAB objects and data types, summary statistics, significance testing and regression analysis, and basic of using MATLAB interactive applications.

WHY USE MATLAB?

In this section, I will discuss the features and capabilities of MATLAB that makes it a suitable tool for the teaching of statistics and data analysis. These are based on my experiences as a MATLAB user who also teaches using MATLAB and this may include comparisons with other packages for data analysis.

Flexibility. MATLAB gives the flexibility of writing commands; therefore, a user can easily create their own functions to perform an analysis that is unavailable.

Linear algebra capabilities. Due to its intuitive mathematical operators, MATLAB is great for linear algebra and allows for operations on vectors and matrices, which is useful when fitting complex models.

Graphical user interface (GUI). Compared to similar programming languages, MATLAB has a friendly graphical user interface and can be quite easy to get around.

Transferability. The programming skills from MATLAB are also transferable and can be a useful if users decide to explore other data science and machine learning specific language such as R and Python (Python Software Foundation, 2018). Table 1 shows an example of a “for” loop in MATLAB, Python, and R. More importantly, it is possible to run R and Python functions from within the MATLAB environment.

Table 1. Creating a loop in MATLAB, R, and Python

MATLAB	R	Python
for i=1:5	for (i in 1:5){	for i in range(1,6):
disp(i^2)	print(i^2)	print(i**2)
end	}	

Vast statistics and machine learning library. MATLAB has built-in functions for statistical data analysis, including summarising data, significance testing, and linear regression. There are also functions for more advance analysis such as generalised linear models, multi-level modelling, cluster analysis, principal component analysis, and supervised machine learning models.

MATLAB apps. The built-in interactive applications give MATLAB a big edge compared to other programming languages because they allow for performing complex analyses using drop-down menu options. After an analysis is completed with the application, this can be converted to code, and the user can edit the code as appropriate. This is a useful tool for learners who can learn by editing aspects of the generated code to modify their analysis. Additionally, one can create interactive applications using MATLAB's App Designer.

MATLAB grader. The grader platform enables assessment of code by specifying an expected outcome and comparing the outcome from student's code to this solution. It is also possible to provide feedback on incorrect answers.

In Table 2, we compare MATLAB to R and SPSS under several themes.

Table 2. A comparison of MATLAB, R, and SPSS

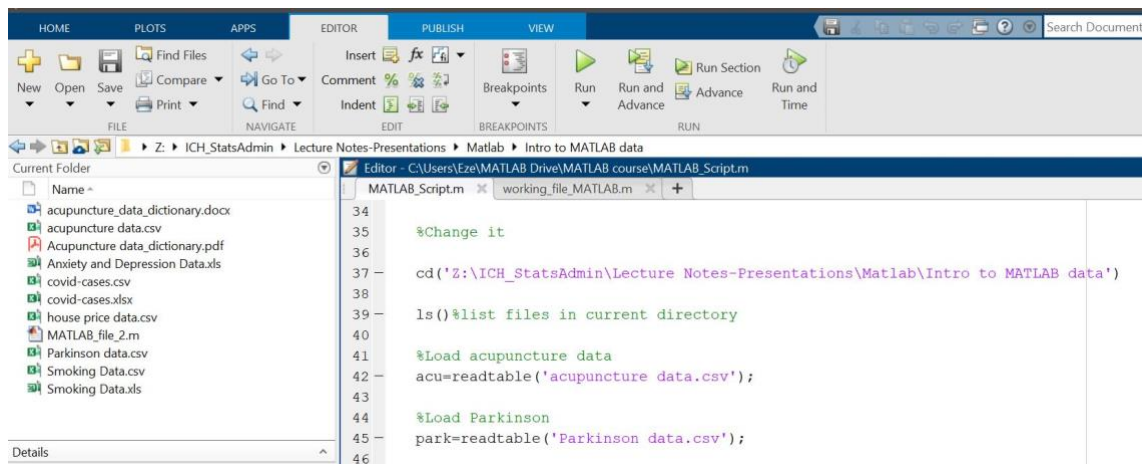
Comparison	MATLAB	R	SPSS
Speed	Fastest for performing numerical computations.	Slower than MATLAB, but faster than SPSS.	Slower than MATLAB and R.
Cost	Requires license (usually available with most academic institutions).	Free to use.	Requires license (usually available with most academic institutions).
Functionality	Useful for engineering applications such as matrix manipulations and signal processing as well as statistical analysis.	Useful for statistical analysis.	Useful for limited statistical analysis.
Ease of learning	Easier to learn for matrix operation and linear algebra than R. Has interactive apps for performing complex analysis from drop-down options. The code used for the analysis can then be generated.	More difficult to learn than MATLAB and SPSS. Matrix operations are not straightforward.	Easiest to learn as all the analysis can be performed using drop-down menu.
Support	MATLAB provides extensive support and a very useful help and documentation page with lot of examples on how to use its different functions.	There are help documentations provided for each base R function. However, some packages have very limited examples.	Extensive help documentation is provided, however examples of usage for these functions are usually not available.

TEACHING STATISTICS WITH MATLAB: USEFUL TOOLS AND TECHNIQUES

In this section, I will describe tools and techniques that I have found useful for teaching data analysis using MATLAB, with an emphasis on online teaching and tools to enhance interactivity. The techniques described here have been applied in the "Introduction to Data Analysis with MATLAB" short course that I developed and currently lead. These tools and technique are broken down into several parts as shown below.

Real data demonstration. Using real data in teaching enables learners to solve real problems and understand the reason for each analysis. This way, students take ownership of their learning and are encouraged to try out techniques learned on their own data. This is a form of authentic learning and similar case studies of authentic learning using statistical software such as R and SPSS can be found in Strangefield (2013) and Rode and Ringel (2019). The dataset used for demonstration should be relevant to the learners and related to their field of interest if possible. The variables used should be clearly explained and learners should not require further reading (other than what is given in the course materials) to understand the data. In our course, three different datasets were made available to the students. Two of these datasets were used for teaching, whereas the third one was only used for the practice exercises.

Code sharing. Actively teaching while sharing code allows learners catch up with course materials and participate in the classroom. This way they can focus more on the context of each chunk of code and why it is needed instead of actively trying to catch up with the teacher's typing. Live code sharing has been implemented in the "Introduction to R" and "Further Topics in R" short courses taught by the Centre for applied statistics courses (CASC). However, this has been done through installing an R package (*Rdrop2*) that connects to Dropbox (see Langan & Wade, 2018 for more details). In MATLAB, code sharing is made possible using MATLAB drive, which is a cloud-based storage for MATLAB files. This drive is useful for using MATLAB online (via a web browser) and through a mobile phone application. For code sharing, MATLAB drive connector is downloaded on the computer used for teaching. This creates a "MATLAB Drive" folder where any script used for live demonstration should be stored. Given that the MATLAB script is now on the cloud, a link to the script can be shared with learners. The link can be set to editable or non-editable. Whenever changes have been made to the script, the teacher needs to save the changes so that it shows up on the web. A Google docs weblink is also provided where learners can share their code. This can be helpful when they get error messages and require support.



```

34
35 %Change it
36
37 cd('Z:\ICH_StatsAdmin\Lecture Notes-Presentations\Matlab\Intro to MATLAB data')
38
39 ls() %list files in current directory
40
41 %Load acupuncture data
42 acu=readtable('acupuncture data.csv');
43
44 %Load Parkinson
45 park=readtable('Parkinson data.csv');
46

```

Figure 1. MATLAB script from live teaching

Live exercises. Having exercises to work through helps learners practice what they have learned and provides assurance that they are comfortable writing their own commands to answer specific questions. At the end of each section of the MATLAB course, an exercise is provided for learners to work through. This is timed, and learners are encouraged to anonymously share their solutions on the Google doc form. Exercises should not be too time consuming because learners may be discouraged from attempting them and then drop off from the rest of the course.

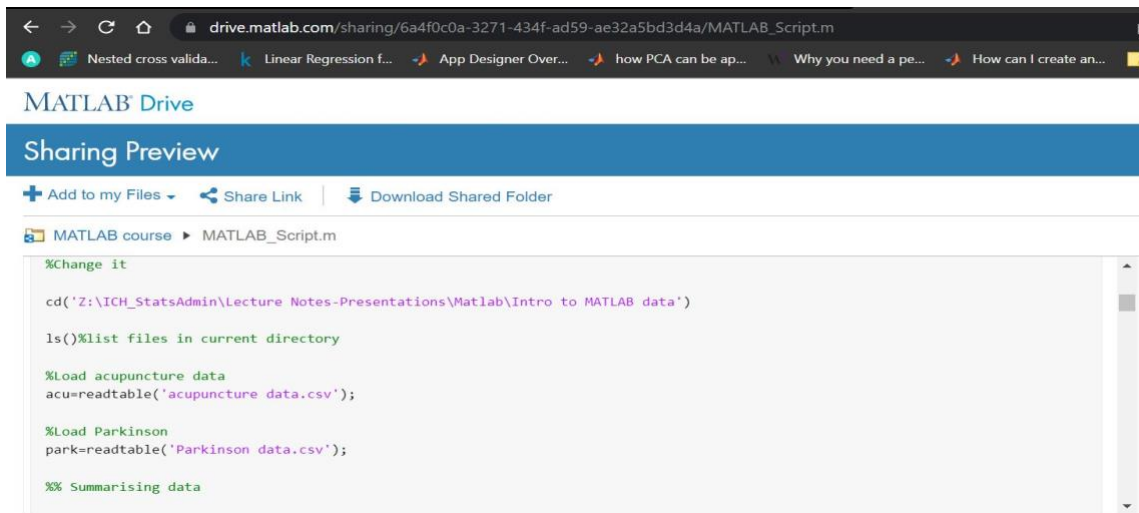


Figure 2: MATLAB live script accessible from the MATLAB drive page

MATLAB grader. This is a useful tool for asynchronous exercises. MATLAB grader can be used on the MathWorks website, or it can be integrated on Moodle via a plugin. This is one way for students to get immediate feedback on the solutions without help from a teaching staff.

TOPICS TO COVER FOR AN INTRODUCTORY COURSE

Table 3 includes a list of topics we cover when teaching statistical data analysis with MATLAB. This is not intended to cover all relevant topics but give a brief overview of them and the MATLAB functions needed in each case.

Topic	Function(s) or Operators	Details
Importing data	readtable()	This covers how to import different data types into MATLAB.
Data subsetting	, &	The AND OR operators can be used the call different parts of a data.
Summarising data	summary(), tabulate(), mean(), median, sd(), iqr()	The summary() function can summarise() each column of a table and requires that variables are correctly defined to get the appropriate summary statistics.
Quantifying Differences	varfun(), groupstats()	Focus on differences in means/medians and differences in proportions between groups. These functions are similar to tapply() in R.
Significance testing	ttest(), ttest2(), chisq()	An introduction to significant tests.
Bootstrapping	bootstrp(), bootci()	Basics of bootstrap sampling and bootstrap confidence intervals.
Linear Regression	fitlm()	An introduction to linear regression.

Table 3: Topics for introduction to data analysis with MATLAB and related functions

STUDENT FEEDBACK

Participant feedback on the MATLAB short course has been mostly positive, with most of them appreciating the focus on using MATLAB for data analysis. This is based on sample of 11 participants who provided some feedback. Some participants have, however, requested to have more time spent on data analysis rather than on data subsetting and manipulation at the start. This feedback was mostly common during the first run of the course, and the pace of the course has been adjusted since then. The feedback also shows that participants in the course have benefited from the live code

web link with the main drawback being that it has to be refreshed each time there is an update on the script.

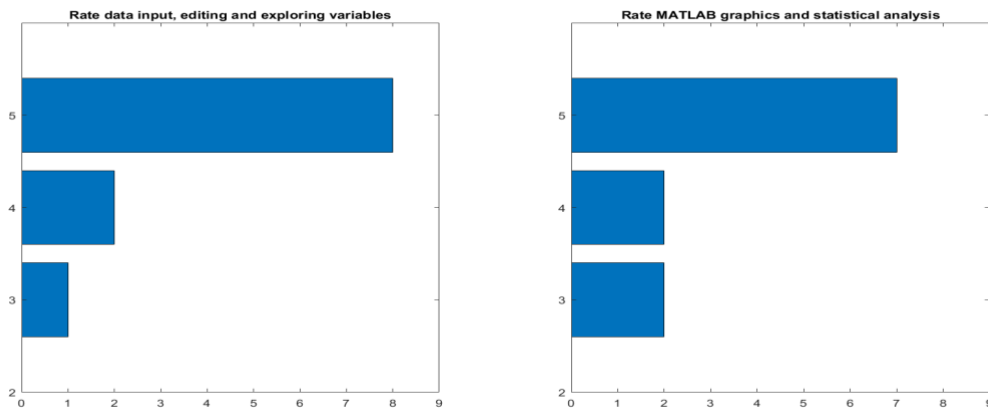


Figure 3. Participant ratings of course from 1 (poor) to 5 (excellent)

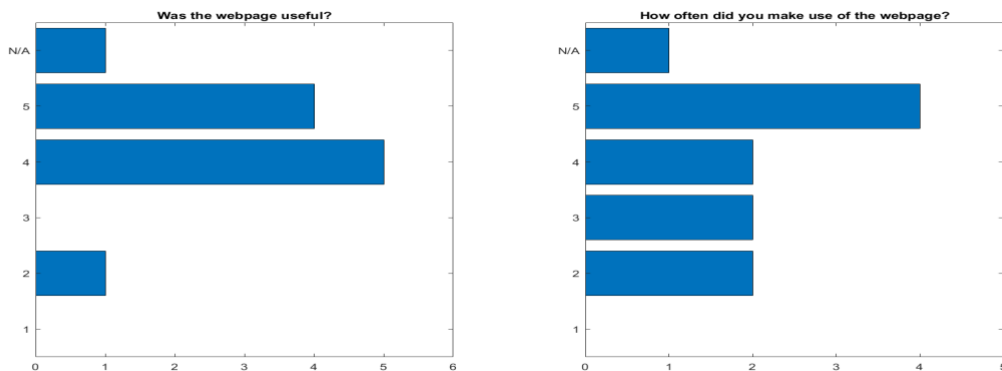


Figure 4. Participant ratings of code sharing webpage. Ratings on usefulness are from 1 (not useful) to 5 (very useful); ratings for frequency of use are from 1 (never) to 5 (very often).

Students were also asked to provide a general comment about the course. Some general comments from the students:

- “I really enjoyed the course and I found it very useful. I would recommend extending it a little bit more every day (maybe 15 min each, 30 min-1h in total) to do more practical exercises.”
- “The instructor was knowledgeable and I understood better how the program fits data.”
- “Chibueze was fantastic - really engaged and engaging lecturer! I hope to go to more of his courses. My only suggestion would be to speed through the basics more quickly (even for Matlab novices, there are easy to grasp), and spend more time on running analyses/showcasing anything that Matlab can do better than the alternatives (any reason to not use R would be welcome!).”
- “I would have spent more time on the actual use of Matlab for statistics and less on Matlab language. Adding more practical sessions with exercises would also be appreciated.”

The above comments suggested that students preferred having more time spent on working through the capabilities of MATLAB for statistical analysis, rather than just the basics of the MATLAB language, which they found easy to grasp. Students also thought more time should be dedicated to practical exercises where they can work on some problems independently.

Finally, we asked students about their suggestions for future courses with the aim of understanding what other knowledge students may find helpful. Some comments are given below:

- “Matlab for machine learning”
- “Intermediate/advanced stats using matlab (after the beginner course)”
- “Next steps in Matlab for statistics?”

- “Programming for psychologist in MATLAB”
- “More complex statistics, like next level stats (after intro), not only 'how to use a software”

The suggestions for future courses were more focused on more advanced statistical analysis using MATLAB. Surprisingly, two students suggested a future course on MATLAB for machine learning, which is not surprising, given the huge interest in machine learning and deep learning in recent years.

CONCLUSION

There is great potential when it comes to teaching statistics with MATLAB, and while MATLAB may require licensing for use, there are free alternatives such as Octave that may not have MATLAB’s vast libraries but are compatible with the MATLAB language. While the course introduced focuses on basic data analysis in MATLAB, the software can be used for more advanced analysis such as machine learning, deep learning, and signal processing. Due to the preference of R and SPSS for data analysis among most of our short course attendees who are usually from a medical background, a MATLAB course may often seem quite niche and not as popular. The MATLAB grader is an excellent tool for student practice and should be used for pre-course activities to provide a blended learning.

REFERENCES

- IBM Corporation. (2021). *IBM SPSS statistics for Windows* (Version 28.0) [Computer software].
- Langan, D., & Wade, A. (2018). Code sharing in statistics workshops. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan*. ISI/IASE. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_9C1.pdf?1531364299
- Python Software Foundation. (2018). *Python language reference* (Version 3.7) [Computer software]. <http://www.python.org>
- R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rode, J., & Ringel, M. (2019). Statistical software output in the classroom: A comparison of R and SPSS. *Teaching of Psychology*, *46*(4), 319–327. <https://doi.org/10.1177/0098628319872605>
- Strange field, A. (2013). Promoting active learning: Student-led data gathering in undergraduate statistics. *Teaching Sociology*, *41*(2), 199–206. <https://doi.org/10.1177/0092055X12472492>
- The Math Works, Inc. (2020). *MATLAB* (Version 2020a) [Computer software].