Coding Practice in Economics: A Survey and Recommendations[1]

Joe Hirschberg[1], Jenny Lye[2]

1 Department of Economics, University of Melbourne, Victoria 3010
Australia.
2 Department of Economics, University of Melbourne, Victoria 3010
Australia.

Abstract

Research in applied economics requires the use of specialised software for data analysis. The scripts or code for these software routines can be written in many ways and although it has recently become standard procedure for these scripts to be submitted along with the research document there has been little attention given to the form of these scripts. In this paper we address this issue by conducting a review of these practices and making recommendations for the creation of better code. Better code is of interest for a number of reasons: It is common practice to use scripts written for one project to be applied in future projects that use similar data and/or techniques of analysis, It has become increasingly common for economic research to be performed by teams that share scripts, thus the need for easy interpretation by different authors, and replication files have increasingly become useful tools for students in graduate courses in research methods.

This paper presents a survey and analysis of code used for 170 papers published in the 2020 volume of the *American Economic Review* and the *Papers and Proceedings* of the 2020 meeting of the American Economic Association that provide code for econometric applications. First, we provide an overview of the most commonly used software combinations for these papers. Then, for the most commonly used software (*Stata*), we examine over 525,000 lines of code and we demonstrate how the style of these scripts can be categorised for their quality. We then provide a style guide for generating the higher quality code with examples for the three most commonly used software packages used for data analysis in applied economic research.

It is now common procedure for academic journals to require the submission of all the code and data used in the generation of the results published in the journal to ensure the replicability of the conclusions presented. This has become especially important with the recent proliferation of journal articles that employ complex estimation procedures applied to large scale datasets with complex structures. In addition, it has become a requirement that data and computer programs be submitted as part of a submission of a research essay or dissertation. [2] Although the policy may be quite specific as to the nature of the data there are few details or guidelines for the code. For example, Vilhuber (2021) uses 5 pages to define the policy for the

[1] This paper is based in part on lectures the authors have given in lectures given to graduate students on research methods and applied econometrics at The University of Melbourne. We wish to thank Lars Vilhuber, Jon Fiva and Dan Millimet, for their comments on earlier drafts.

[2] In this paper we will refer to programs, scripts and code interchangeably to mean a computer readable file that lists a series of instructions to be read by software designed to carry out data manipulations and statistical computations.

American Economic Association publications but scant attention to the condition of the code submitted.[3] However, aside from assuming the program can be run to produce the results in the publication, there is no condition as to the "readability" of the code to ensure that it does what it proports to do. As we demonstrate below, the code is written in a manner that it is difficult to follow thus, although it may generate the reported results it may be difficult to follow or adapt for an alternative application. One way to avoid these pitfalls is to write code that is clearly written and uses a clear style.[4]

There are few occasions when a computer code work or perform the task perfectly the first time they are run. This implies that they need to be debugged. A well written program is easier to debug and maintain and is more useful to others who may want to replicate your work, extend it, to speed it up, or borrow from it.[5] Good style helps the reader to concentrate on parts of code elements that can be checked separately. Additionally, the availability of well written code has become a useful tool in the training of future researchers.

Another factor in the need for good code style is the rise of "Research Teams" in economics. Recently Jones (2021) has documented the rise of multiple author contributions in economics over solo author papers. When multiple authors work on a project it is common for different authors to share code to be used by others. This may necessitate that each element in the overall project be available for all the team members to contribute. It may also be the case that multiple projects use a common set of datasets or algorithms which requires that code be in a form that allows it to be reused for different projects.

For most applications, the construction of code involves assembling data into a form that can be used by computation programs (e.g. regressions) that have been written by professional programmers. Programs have been written to invert matrices, multiply matrices, find the eigenvalues of matrices, etc. are widely available. The algorithms for these computations have been the subject of intense investigation by many authors over the past 70+ years.[6] Many statistical packages have optional elements that allow the researcher to design a special purpose estimation procedure, however the focus in this paper is the construction of code to manipulate data in a transparent way so that an analysis can be replicated and assumptions used made clear. We will draw on statistical packages such as *Stata*, *SAS* and *R* as examples.

We have conducted a survey of current programming practice by examining the *Stata* code submitted for the replication of analysis conducted for the papers published in volume 110 (2020) of the *American Economic Review* and the Papers and Proceedings of the 2020 meeting of the American Economic Association. In this survey we also categorise code as their similarity to two diametrically different sets of code that perform the same tasks.

[3] L. Vilhuber (2021), "AEA Data and Code Availability Policy", *AEA Papers and Proceedings,* 111, 818-823. https://www.aeaweb.org/articles?id=10.1257/pandp.111.818 . For more information see: https://www.aeaweb.org/journals/data/faq#package .

[4] The code we discuss here is in the nature of "scratch programs" that intended to be written in high level languages such as *Stata, R* and *SAS* and may only be used a limited number of times. Production programs that are used repeatedly to perform large scale tasks require more detailed specifications that are outside the scope this study.

[5] Many of these details reflect longstanding and general advice (e.g., Kernighan and Plauger 1978).

[6] See Press, W., S. Teukolsky, W. Vetterling, B. Flannery, and M. Metcalf, (2007), *Numerical Recipes: The Art of Scientific Computing,* 3rd ed, Cambridge University Press, for details on many of these algorithms.

The paper proceeds as follows: first, we present the results of a computer generated survey of the code provided for the 2020 volume of the *American Economic Review* and the *Papers and Proceedings* of the 2020 meeting of the American Economic Association. Then we provide a style guide for some of the best practice in writing code with attention to the main elements of the code, and then we discuss examples in *Stata*, *SAS* and *R Stata*, *SAS* and *R* code for performing some of the most common tasks

Key words: *Stata*, *R*, *SAS*, Replication, analysis of code

JEL: C8, Y1, A23, C55, L86