Can R Notebooks help with reproducibility?

Kevin Ha Ola Andre Olofsson

Introduction

In this document we will look at reproducibility and the importance of it. In this context, we will address the topic of using "R Notebook" in RStudio. We'll look into theory regarding reproducibility and R Notebook, which we will also follow up with a discussion.

What is reproducibility?

In science it's by some considered to be the holy grail of statistics. Basically it means doing an experiment or research multiple times by different people or groups, with the same numbers and get the same result. This is desirable because it is important that the findings are robust, trustworthy, and conducted in a satisfactory way.

Researchers often use the terms replicability and reproducibility interchangeably, but it is useful to distinguish between them. Replicability is "re-performing the experiment and collecting new data," whereas reproducibility is "re-performing the same analysis with the same code using a different analyst" Patil, Peng, and Leek (2016). According to Stanovich (2014), one of the most important criteria for scientists is that the findings are presented in a way that they can be replicated, exposed for criticism, or extended further on.

Challenges

One of the big challenges of Data science has been to reproduce the identical results from an earlier research. In a significant amount of research it has proved hard or flat out impossible to reproduce the results from other peoples and older research for a multitude of reasons. One of the big reason is quite simply that the data used in the research is unobtainable, due to them being "too old" or lost. Another reason is the author doesn't want to share the data or can't find it (again, lost). Other challenges is the usage of different systems, equipment, and so on. Theories around the research and calculations can also change over time.

Computable Documents

In more modern times a new idea around this have been brought up: computable documents. The idea behind this is when publishing a research paper, one submits the data, equations and

calculations together with a document that's computable. That way when other researches wants to run the experiment, they have access to everything, and can see what others did before them.

Computable Document Format (CDF) is an electronic document format designed to allow authoring dynamically generated, interactive content "Computable Document Format" (2021). The format was designed to emprove or kill the PDF by making a document the reader could interact with, by using sliders, menus, and buttons which PDF does not allow. This consequently increases engagement in the readers and their understanding.

So the question then becomes: can computable documents like R Notebook help with reproducibility?

Short literature review

According to McNutt (2014) advancing in science is reliant on discoveries that can be trusted, but stated that studies that have been performed, can't be reproduced. In tandum with this, Peng Peng (2011) has brought up the idea that reproducibility should be a requirement for publishing research.

Already in the early 1930's the accessability to the data for a research was aired by Frisch He stated that in statistical research, the raw data should be published Frisch (1933). Later in the 1960's it was deemed nearly impossible to reproduce research on economy with big models. Then in 1982 there was a research project done: the Journal of Money, Credit and Banking, where they tried to replicate research article with only the submitted data. This is looking back, very close to reproducibility in modern terms. The results was a staggering 2 of 70 articles could be reproduced. They concluded was because for the most part missing data, documentation and computer systems, as quite few of the authors would provide their data and coding Dewald, Thursby, and Anderson (1986). To counteract this, they came up the solution to store the coding and raw data used in research articles, which in a sense is the foundation of computable documents today.

Discussion

Reproducibility is one important approach that scientists use to gain confidence in their conclusions McNutt (2014). How confident in a conclusion would one be if the scientist could not reproduce the result and conclude similarly again? She states that this confidence is important regarding the broad scientific community. This is due to the scientifical knowledge is public in a special sense. It does not only exist in the mind of the particular researcher, but one could argue that the knowledge does not exist until it has been submitted to the scientific community for criticism and empirical testing Stanovich (2014). He elucidates this matter by introducing the term replication. The term is understood to mean that a finding must be presented to the scientific community in a way that allows fellow scientists to re-do the study and achieve the same results.

According to a analysis conducted by McCullough (2009), most economics journals take no substantive measures to ensure that the results they publish are replicable. Top economics journals have been adopting mandatory data+code archives in the past few years. The movement toward mandatory data+code archives has yet to reach the open access journals. Open Access (OA) journals are often perceived, rightly or wrongly, as having a second-class status compared to traditional journals. Note that in the list of top 50 journals in Table 1, not a single journal is OA.

Some important points to highlight that can cause problems with reproducibility in R Notebook is:

1. System or version incompatibility

- Packages, software, etc.

2. Advancement or changes in theories
3. Human error

- May only be reduced to a certain degree
- Forgetting to add information and data

4. Motivation

- Certainty relies on willingness
- Willingness to turn in all data, code, information, etc.

Firstly, there are different OS and OS platforms, which can create problems. Software and packages being updated can cause incompatibility and across different versions. To fix this one needs a huge systematic archive of downloadable software and packages as it was at the time. Another issue is that a theory about a subject today can change in 20 – 30 years, which can affect how we go about it and in worst case, calculations and formulas can change.

There's also the human factor. Human error such as a misunderstanding of the subject or a typo and so on can create follow up mistakes through the document. In huge documents with 1000s of observations and multiple formulas and calculations can make a small mistake into a big one that can be hard to track and fix, which makes reproducibility borderline impossible.

Finally, there's also willingness. Creating a big R Notebook or computable document with 1000s of observations, formulas, calculations, analysis, creating data sets and sourcing it all can be a huge workload. If the person or group takes a shortcut or leaves something out because they think it's not necessary to write down or forgets to add it, we're once again missing information. Without that information we can't reproduce the research, and if we must ask the author for the data and information they didn't add, part of the point with

computable documents for reproducing is lost. On top of that, we're relying on the author(s) actually having the data on hand or remembering it correctly after potentially years.

Why is reproducibilty important? What are the issues surrounding it? What steps have been taken to improve the situation? What does scientist say about it?

Conclusion

The question was «can R Notebook help with reproducibility?». The short answer is yes, it can help, but it doesn't solve the problem, atleast not yet. There's quite a few issues with the current R Notebook system and computable documents.

To summarize, computable documents like R Notebook definitly helps with reproducibilty, but it doesn't solve the issue, atleast not yet. There are currently too many problems both human and technical. There's progress in these areas like using identical computer setups from a server, libraries of software and packages, as they were, are currently being developed, the option to check the system for the setup it used when the research was first done, and so on. So while it doesn't fix the issue, it jelps and it's definitly a step in the right direction.

Reference

"Computable Document Format." 2021. *Wikipedia*, July.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *The American Economic Review* 76 (4): 587–603.

Frisch, Ragnar. 1933. "Editor's Note." *Econometrica* 1 (1): 1–4.

McCullough, B. D. 2009. "Open Access Economics Journals and the Market for Reproducible Economic Research." *Economic Analysis and Policy* 39 (1): 117–26. https://doi.org/10.1016/S0313-5926(09)50047-1.

McNutt, Marcia. 2014. "Reproducibility." *Science* 343 (6168): 229–29. https://doi.org/10.1126/science.1250475.

Patil, Prasad, Roger D. Peng, and Jeffrey T. Leek. 2016. "A Statistical Definition for Reproducibility and Replicability." Cold Spring Harbor Laboratory. https://doi.org/10.1101/066803.

Peng, Roger D. 2011. "Reproducible Research in Computational Science." *Science* 334 (6060): 1226–27. https://doi.org/10.1126/science.1213847.

Stanovich, Keith E. 2014. *How to Think Straight about Psychology*. Ninth edition, Pearson new international edition. Harlow, Essex: Pearson.

Appendix