Can R Notebooks help with reproducibility?

Introduction

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.

Challenges

One of the big challenges of (statistical?/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. 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 (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 (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. (Kilde. Arnstein sin e paywall) 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

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.

Just to name a few there's different OS and OS platforms that can create problems. Software and packages being updated can cause incompatibility and across different versions. To fix this one need 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 og 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 & calculations can make a small mistake into a big one that can be hard to track and fix, which makes reproducibility borderline impossible.

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 have to 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) actualy having the data on hand or remembering it correctly after potenially years.

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