Can R Notebook help with reproducibility?

Introduction

This text will focus on how R Notebook can help with reproducibility. The text is based on reproducibility as a standard for general research being published, as well as reproducibility in the field of economical studies. R Notebook will be discussed as a solution to meet the criteria of reproducibility.

Literature review

The following articles explained in this part is to be used in the discussion:

1. *"Reproducible Research in Computational Science"*

2. *"Do economics journal archives promote replicable research?"*

3. *"Statistical Analysis and Reproducible Research"*

"Reproducible Research in Computational Science," by Peng (2011).

Roger D. Peng presents reproducibility as the bare minimum standard when evaluating all published scientific findings. The reasoning being the change in technology leading to large public data sets used by multiple scientist to produce independent research. This change making it harder to investigate the credibility of different studies/papers.

According to Peng, the studies/papers needs to have data and code public to make it possible for it to be reproducible. An issue following this requirement is that the computer code often is not available due to software systems and packages being private.

As the main goal presented in the article is to develop a culture that require reproducibility for computational science to be publish, two solutions are mentioned:

1. Authors make their data and code available for everyone by using the free code repositories. Such as Github and SourceForge.

   - This would make it possible to provide a minimum code as online material which would be informative, and could be used to reveal any potential problems.
   - The next step would be to be able to provide datasets in durable non-proprietary format, but this would require additional costs.

2. The scientific community creates a single place for authors in all fields to make their studies reproducible.

   - In order to accomplish this the community would need to create a common DataMed central, CodeMed central and PubMed central where publishers data, metadata, and code can be stored with their studies. This would require additional costs and would need to be coordinated and supported by the government.

The article makes it clear that even if a analysis is reproducible it does not guarantee validity, quality and correctness even when it is published to others.

"Do economics journal archives promote replicable research?" by McCullough, McGeary, and Harrison (2008).

McCullough's article puts focus on reproducibility and replication in the field of economical studies. In this article the ability to replicate a study is put as the gold standard to evaluate the reliability of a scientific claim, and should be aimed for in all archives. Replicability meaning being able to collect new data and get the same results/conclusions.

The article shines light on the fact that long-standing archives of economics do not accustom to reproducibility. And some of those who claim to be data-only archives fail in reality due to authors being in charge of submitting their code and data, but neglecting to do so. In these cases the papers get published regardless as a result of lack of controlling by the journal archives. Meaning there is improvement to be done to the promotion of replicable research.

Additionally, the article shows that editors will defend published work even if it is not replicable with the argument of the papers being published before the requirements for data availability. This only showcases the importance of stricter data+code requirements.

"Statistical Analysis and Reproducible Research," by Gentleman and Lang (2004).

In the discussion on how R Notebook can help meet the criteria of reproducibility Gentleman and Lang's definitions on compendium, code chunks and text chunks are used.

> *Compendium is both a container for the different elements that make up the document and its computations (i.e. text, code, data, . . .), and as a means for distributing, managing and updating the collection.*
>
> *Code chunks are sequences of commands in some programming language such as R or Perl. Code chunks are intended to be evaluated according to the language in which they are written. These perform the computations needed to produce the appropriate output within the paper, and also to produce intermediate results used across different code chunks.*
>
> *Text chunks describe the problem, the code, the results and often their interpretation. Text chunks are intended to be formatted for reading*

Discussion

This discussion contains three main parts based on the literature review:

1. *Can R Notebook solve the problems surrounding reproducibility?*

2. *An example of "Code chunks" and "Text chunks" in R Notebook.*

3. *Do authors take initiative to publish reproducible papers or do they need to be forced?*

4. *Will the current focus and interest on reproducibility remain or will it decrease in the future?*

Can R Notebook solve the problems surrounding reproducibility?

To meet the criteria of reproducibility computable documents is a solution. R Notebook is one of the available programs.

R Notebook have the following features:

1. Uses R markdown, collects all parts into one document
2. Codes goes through sequentially, but can also go through individually
3. Uses Latex - making it possible to write mathematics
4. Uses pandoc as a transformer - html. pdf. and word.

These feature is what makes the program a computable document as a working compendium, with working code and text chunks and the ability to transform. A computable document meaning the container for the different elements that make up a document and its computations (Gentleman and Lang 2004). These code chunks being sequences of commands in the programming language R (Gentleman and Lang 2004). With the possibility to the describe the code, the results and their interpretation with text chunks (Gentleman and Lang 2004).

With the listed features R Notebook is a tool that can be used to be able to produce reproducible papers/studies. It is still important to keep in mind that one can not guarantee that each user of R Notebook creates reproducible research, as the program does not guarantee for human mistakes not to be made.

To see R Notebook in context with Peng's "Reproducible Research in Computial Science" article, R Notebook can be used as a step to reach the goal of developing a culture that requires reproducibility in computational science. This by giving authors a system to provide data and code for everyone as it connects to free repositories. This proves that R Notebook is a great tool to reach the goals presented in the article, even if it can not guarantee for every paper/study using the program being reproducible.

.

An example of "Code chunks" and "Text chunks" in R Notebook:

To showcase R Notebook working as a compendium it can be demonstrated by presenting how code and text chunks works together. Everything written in this document would be a text chunk and the following chunk would be a code chunk:

```
sessionInfo()
```

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.1.1    magrittr_2.0.1    tools_4.1.1       htmltools_0.5.1.1
##  [5] yaml_2.2.1        stringi_1.7.3     rmarkdown_2.10    knitr_1.33
##  [9] stringr_1.4.0     xfun_0.25         digest_0.6.27     rlang_0.4.11
## [13] evaluate_0.14
```

This code chunk is put in the document to show which packages are being used in R Notebook as it is being written.

In more advanced papers code chunks would show how the data is processed and the text chunks would be used to explain the results.

These documents being able to transform to word., pdf, and html. is what makes the compendium a computable document.

.

Do authors take initiative to publish reproducible papers or do they need to be forced?

Even as the previous questions shows that R Notebook can be used as a tool to best publish reproducible papers, McCullough's paper "Do economics journal archives promote replicable research?" proves that some authors might need stricter requirements to follow it through.

Many archives do not promote reproducibility, as it is proven that some authors do not submit their data and code but their papers are still getting published (McCullough, McGeary, and Harrison 2008). The article also brought up the problem regarding some editors in the archives supporting articles that are not replicable if the authors published it before requirements for data availability.

This article presents an issue in the field of economical studies regarding publishing reproducible and replicable articles. Therefor, it should be set stricter requirements forcing the authors to publish reproducible papers as there is not enough authors taking the initiative to do so.

.

Will the current focus and interest on reproducibility remain or will it decrease in the future?

It is valid to question whether the current interest on reproducibility will remain in the future or if it will decrease. As mentioned in Rogers Peng's article "Reproducible Research in Computational Science," reproducibility is put as the bare minimum standard when evaluating published scientific findings. Reproducibility might not be enough as a minimum standard moving forward, as the technology further develops. It is likely to require more of the published scientific findings as Peng's article states that reproducibility still does not guarantee quality.

In McCullough's article "Do economics journal archives promote replicable research?" the lack of routines and control surrounding the standard of reproducibility and replicability is presented. The article indicates that there is a need for stricter requirements surrounding published scientific papers. As reproducibility has been the bare minimum standard for a while and there is still studies not meeting the requirement, the focus and interest on reproducibility will likely remain. When the standard is met in all published papers a higher minimum standard can be devloped.

Conclusion

To conclude R Notebook is a computable document that has the features to be used as a tool to help with reproducibility. It gives authors support to publish reproducible research, but it can not guarantee it due to the possibility for human errors. For authors to be motivated to use the program, it would be efficient to have stricter data and code requirements in publishing archives. This would lead to the authors being forced to use programs such as R Notebook more actively and would be a help to the issue of reproducibility.

References

Gentleman, Robert, and Duncan Temple Lang. 2004. "Statistical Analyses and Reproducible Research." *Bioconductor Project Working Papers*, May. https://biostats.bepress.com/bioconductor/paper2.

McCullough, B. D., Kerry Anne McGeary, and Teresa D. Harrison. 2008. "Do Economics Journal Archives Promote Replicable Research?" *Canadian Journal of Economics/Revue Canadienne d'économique* 41 (4): 1406–20. https://doi.org/10.1111/j.1540-5982.2008.00509.x.

Peng, Roger D. 2011. "Reproducible Research in Computational Science." *Science* 334 (6060): 12261227. https://doi.org/10.1126/science.1213847.

Appendix