

Can R Notebooks help with reproducibility?

Assignment 1 - MSB 105 - Ole Alexander Bakkevik & Sindre Espedal

Introduction

One might argue that there are two basic reasons to be concerned about making research reproducible.

The first is to show evidence of the correctness of your results. Descriptions contained in scholarly publications are rarely sufficient to convince skeptical readers of the reliability of our work. In simpler times, scholarly publications showed the reader most of the work involved in getting the result. The reader could make an informed choice about the credibility of the science. Now, the reader may feel they are being asked to blindly trust in all the details that were not described in the original journal article.

Adopting a reproducible workflow means providing our audience with the code and data that demonstrates the decisions we made as we generated our results. This makes it easier for others to satisfy themselves that our results are reliable (or not, since reproducibility is no guarantee of correctness).

The second reason to aspire to reproducibility is to enable others to make use of methods and results. Equipped with only our published article, our colleagues might struggle to reconstruct our method in enough detail to apply it to their own data. Adopting a reproducible workflow means publishing our code and data in order to allow scientists to extend our approach to new applications with a minimum of effort. This has the potential to save a great deal of time in transmitting knowledge to future researchers. *Reproducibility Guide* (n.d.)

In this paper we will discuss the topics mentioned above.

Literature review

Reproducibility, R notebooks

Peng (2011) states that “The standard of reproducibility calls for the data and the computer code used to analyze the data be made available to others.” As a standard, this creates a tedious and non-effective approach to replication. A far more beneficial process is to independently inspect utilized data variables. R-notebooks and other reproducible systems could serve as a crucial component in verifying scientific results.

Replicability

Being able to replicate research results by other researchers is one important part of the methodology in science. In the past, there has been little testing of replicability. Reasons for this are that it is not promoted to replicate another researcher's work. Criticism can also arise about lack of creativity and imagination. A critical question is also asked to the integrity of the researcher as one can be interpreted as critical to the findings or that one does not trust the researcher. Such arguments makes it less attractive to conduct replication studies.

Dewald et al. (1986) tried to replicate a number of data-sets and they found that accidental errors in empirical articles are rather more common than unusual. Although it is quite common for errors to occur in empirical economic research, it is quite frustrating and difficult to replicate and build on the research when there are many errors in the data-sets. This does not appear to significantly affect the conclusion of the authors.

In recent times, technology has made it easier, cheaper and more efficient to make and maintain journal archives. Still McCullough et al. (2008) finds that the potential offered is reduced when editors fail to enforce and authors do not adhere to the guidelines of the journal archives. It is noted that few researchers use the opportunity as offered to engage in replication because economic profession is considering replication as an ideal “to be known but not to be practiced” (McCullough et al., 2008).

Possible solutions

Compendium and “Code Chunks”

Gentleman and Lang (2007) points out that a computable compendium might be an important tool for integrating codes and data etc. This is because when such tools are collected and assembled it must be possible to distribute and update, given that the compendium is of the right quality, so will the possibility of reproduction be simple.

Another possible solution is “Code Chunks” or “Text Chunks.” Code and text chunks are a tool used to display data and code for illustrations. Text chunks are used to describe and interpret results and codes. Dynamic document will hence be an optimal compendium since all the data and components will be available for reproduction (Gentleman and Lang, 2007).

Incentivizing Reproducibility

Over the past several years, a series of publications and policy statements have generated increasing awareness in the scientific community of the scale and implications of the problem of irreproducible data—or at least lack of robust results—particularly in the realm of basic and translational research.

Recent studies have shown that the key findings in 50% or more of published reports in certain fields cannot be reproduced. As the public, government, and private funders of research comprehend the extent of the problem, trust in the scientific enterprise erodes, and confidence in the ability of the scientific community to address this problem wanes. In addition, there is considerable potential for reputational damage to scientists, universities, and entire fields (for example, cancer biology, genomics, and psychology). (*An Incentive-Based Approach for Improving Data Reproducibility*, n.d.)

One possible cause of irreproducible-data is stated by Hessen as “*Scientists are incentivized to produce more results at the expense of spending more time on the reproducibility of any given result.*” Hessen furthermore list three possible solutions:

- One solution is to eliminate imperfections in the peer review system.
(*Without those imperfections credit incentives are perfectly aligned with the social optimum in Hessen’s model*)
- Another solution focuses on the amount of credit given for irreproducible results.
(*If the credit given to irreproducible results matched the social value of those results more closely, the gap between the credit-maximizing optimum and the social optimum would be reduced*)
- A third solution aims to compensate for the misalignment.
(*limiting the number of papers scientists may publish per unit time*) (Schulz et al., 2016)

Incentivizing gone wrong

A good example of fraudulent science is Andrew Wakefield and his study on the link between autism and the MMR vaccine published in the Lancet. Wakefield was paid by a Legal Aid Board of parents of children with autism to conduct a pilot study of virological investigation in autistic children, some of whom were included in the Lancet publication. Additionally, Wakefield most likely manipulated the data, thus presenting false results. Since then Wakefield has become the “*godfather*” for the anti-vaccine movement, a movement whom have grown exponentially during the covid-19 pandemic. (Schulz et al., 2016)

Example list 2 level

Example of a *code chunk* in an R Notebook:

```
l <- list(x = 1:4, y = c(TRUE, FALSE, FALSE), z = c("aa", "bb"), zz= c(2.1, 4.33))
str(l)
```

```
## List of 4
## $ x : int [1:4] 1 2 3 4
## $ y : logi [1:3] TRUE FALSE FALSE
## $ z : chr [1:2] "aa" "bb"
## $ zz: num [1:2] 2.1 4.33
```

Session info

Example session info:

```
sessionInfo()
```

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Catalina 10.15.7
##
## 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
```

The session info function provides the reader information regarding which operating system, packages and data sets that have been used. This information is crucial in terms of gaining reproducibility.

Reproducibility across sectors

Other areas where application of reproducibility would prove beneficiary is e.g. the pharmaceutical industry. Present day studies show that replicating present day clinical-research data is demanding. Which often leads to drugs having to prolong their release to actual patient trials. One human factor could be the fear of being “discredited” among peers, which lead to an bias among researchers. Ultimately causing studies not to be reproduced. (*Why Is Reproducing Pharmaceutical Medical Research so Hard?*, n.d.)

Conclusion

Providing studies that are reproducible is vital in terms of quality assurance and cost- effectiveness. In addition deterring fraudulent scientists is crucial.

By using dynamic documents in the form of codes, data, explanations, etc. in the form of code chunks and text chunks, there are good opportunities for both replication and reproducibility of research, and also further research on previous studies.

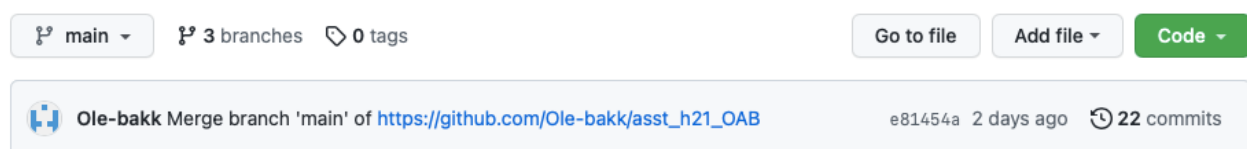
- Motivate researchers to share to make their work available
- Disadvantage? maybe too many different packages (difficult to keep track)

References

- An incentive-based approach for improving data reproducibility.* (n.d.). <https://www.science.org/doi/full/10.1126/scitranslmed.aaf5003>
- Dewald, W. G., Thursby, J. G., and Anderson, R. G. (1986). Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. *The American Economic Review*, 76(4), 587–603.
- Gentleman, R., and Lang, D. T. (2007). Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, 16(1), 1–23. <https://doi.org/10.1198/106186007X178663>
- McCullough, B. D., McGeary, K. A., and Harrison, T. D. (2008). Do economics journal archives promote replicable research? *Canadian Journal of Economics/Revue Canadienne d'économique*, 41(4), 1406–1420. <https://doi.org/10.1111/j.1540-5982.2008.00509.x>
- Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, 334(6060), 1226–1227. <https://doi.org/10.1126/science.1213847>
- Reproducibility guide.* (n.d.). <https://ropensci.github.io/reproducibility-guide/sections/introduction/>
- Schulz, J. B., Cookson, M. R., and Hausmann, L. (2016). The impact of fraudulent and irreproducible data to the translational research crisis solutions and implementation. *Journal of Neurochemistry*, 139(S2), 253–270. <https://doi.org/10.1111/jnc.13844>
- Why is reproducing pharmaceutical medical research so hard?* (n.d.). <https://www.pharmaceutical-technology.com/features/why-is-it-so-hard-to-reproduce-medical-research-results/>

Appendix

Display of Git commits and three branches



The screenshot shows a GitHub repository interface. At the top, there are three buttons: 'main' (selected), '3 branches', and '0 tags'. To the right are 'Go to file', 'Add file', and 'Code' buttons. Below this, a commit message is displayed: 'Ole-bakk Merge branch 'main' of https://github.com/Ole-bakk/asst_h21_OAB'. To the right of the message is the commit hash 'e81454a' and the text '2 days ago'. At the bottom right, it says '22 commits'.