«R Notebooks» og reproduserbarhet

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

disposisjon (dere trenger ikke dekke alt listet her)

Introduction

There is mainly 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 sceptical 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* (u.å.)

In this paper will discuss the topics mentioned above in an light,yet hopefully understandable manner.

Reproducibility, R notebooks

Roger D. Peng states in his article "Reproducible Research in Computational Science, 2011" that "The standard of reproducibility calls for the data and the computer code used to analyze the data be made available to others" Peng (2011).
As a standard , it creates a tedious and non-effective approach to replication. A far more beneficial process is to independently inspect utilized data variables. R-notebook and other reproducible systems would serve as an crucial component in verifying scientific results.

- Litteraturgjennomgang

*** Replikerbarhet/reproduserbarhet**

*** Problemets omfang**

- **Vil dagens løsning med arkiv av data og event. programkode hos**

tidsskriftene kunne løse problemet?

* Mulig løsning (teoretisk plan):

- «Compendium», «Dynamic document», «code chunck» og «text chunk»

* Mulig løsning:

- R Notebooks

- Analyse

* Løser R notebooks problemet med reproduserbarhet

- helt eller bare delvis

* Eksempler på «code chunks» («R Code Block») og «text chunck» i R notebook

* Har forskerne incentiver til å være «reproduserbare», eller må de tvinges?

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* (u.å.)

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.
  (l*imiting 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 representing in 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

```
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 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  fastmap_1.1.0   tools_4.1.1
##  [5] htmltools_0.5.2 yaml_2.2.1      stringi_1.7.4   rmarkdown_2.10
##  [9] knitr_1.33      stringr_1.4.0   xfun_0.25       digest_0.6.27
## [13] rlang_0.4.11    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 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?* (u.å.)

Conclusion

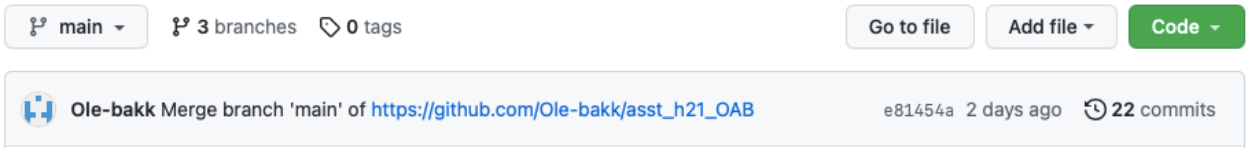
Providing studies that are reproducible is vital in terms of quality assurance , cost- effective and deterring fraudulent scientists is crucial.

For generelle tanker rundt reproduserbarhet er Peng (2011) en god kilde. Videre gir McCullough et al. (2008) en god illustrasjon av problemets omfang innen fagområdet økonomi. McCullough et al. (2008) diskuterer også om tidsskriftenes arkiver av datasett og programkode er en tilfredsstillende løsning av problemet

Litteraturliste

Appendix

Display of Git commits and three branches



An incentive-based approach for improving data reproducibility. (u.å.). https://www.science.org/doi/full/10.1126/scitranslmed.aaf5003

McCullough, B. D., McGeary, K. A., og 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. (u.å.). https://ropensci.github.io/reproducibility-guide/sections/introduction/

Schulz, J. B., Cookson, M. R., og 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? (u.å.). https://www.pharmaceutical-technology.com/features/why-is-it-so-hard-to-reproduce-medical-research-results/