Reproducible Research with RMarkdown

Alexandra Posekany

SS 2020

Reproducibility in science

The Oxford Dictionaries Online defines the scientific method as "a method or procedure that has characterized natural science since the 17th century, consisting in systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses". One of the most important properties that a method of scientific inquiry must satisfy is: reproducibility! The practice of experimental control and reproducibility can have the effect of diminishing the potentially harmful effects of circumstance, and to a degree, personal bias.

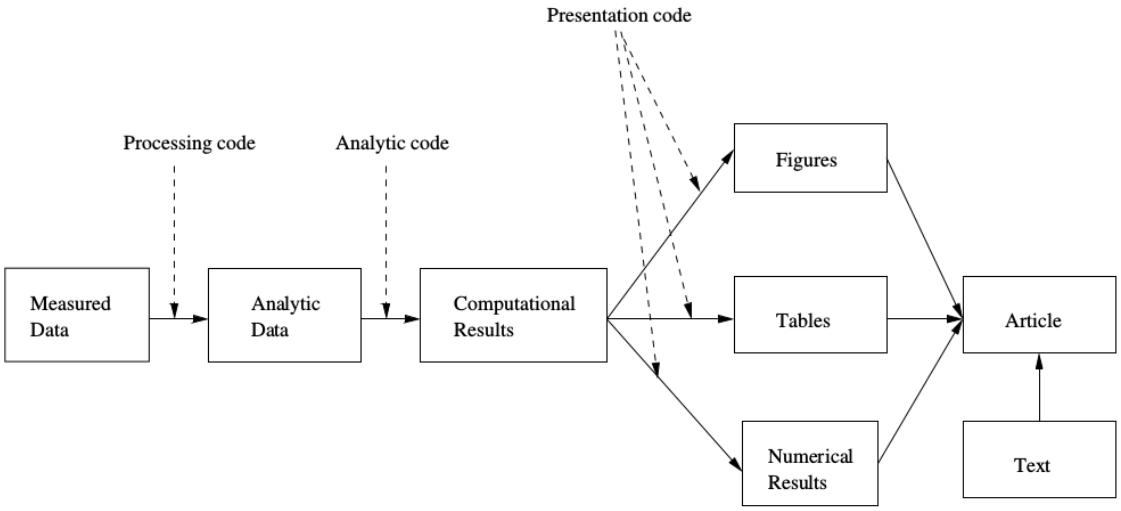
According to a 2016 poll of 1,500 scientists reported in the journal Nature, 70% of them had failed to reproduce at least one other scientist's experiment (50% had failed to reproduce one of their own experiments). This raises the question how to make research more reproducible. According to Prof. Roger Peng, Department of Biostatistics and John Hopkins Bloomberg School of Public Health (Peng 2009) it has to fulfil the following criteria:

- Analytic data must be available
- Analytic code must be available
- Documentation of code and data
- Standard means of distribution

The data analysis workflow

RMarkdown can aid in increasing the reproduciblity of data driven research, since it

- Maps the entire process of statistical research in one document!
- generates dynamic report, which can be updated automatically if data or analysis change (Friedrich Leisch)



RMarkdown: How it works.

RMarkdown provides an unified authoring framework for data science, combining your code, its results, and your prose commentary. It, thus, combines,

- Knitr (Xie 2017) (a further development of Sweave (Leisch 2002)) and
- Pandoc Markdown (a slightly revised version of the markup language Markdown (by John Gruber) which can handle multiple output formats and has added new functionalities)
- which has been well integrated into the RStudio IDE.

R Markdown documents are fully reproducible and support dozens of output formats, like PDFs, Word files, slideshows, and more.

A good documentation of Rmarkdown can be found under

- http://rmarkdown.rstudio.com
- http://rmarkdown.rstudio.com/gallery.html

Rmarkdown documents consist of three different parts

- a YAML header
- Markdown statements
- knitr R chunks

Let us first have a look at a simple example and then study these three elements separately.

```
---
title: "The Iris Data Set"
author: "Alexandra Posekany"
date: "SS 2020"
output:
  pdf_document: default
  html_document:
    df_print: paged
bibliography: bibliography.bib
---

```{r setup, include=FALSE}
library("xtable")
library("stargazer")
data("iris")
```

***

Description

This famous dataset [@fisher, @anderson] gives the measurements in centimeters of the variables sepal l

![Illustration of the Variables of the iris data set.](iris.png)
```

```
Iris is a data frame with 150 cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Pe

```{r summary1, echo = FALSE}
#summary(iris)
```

```{r echo = FALSE, results = "asis"}
options(xtable.comment = FALSE)
xtable::xtable(summary(iris), type = "latex", comment = FALSE, caption = "Summary of the data set")

```

Scatterplot Matrix

```{r scatterplot, echo = FALSE}
plot(iris)
```

Logistic Regression Analysis

```{r echo = FALSE, include = FALSE}
iris[['Is.Versicolor']] <- as.numeric(iris[['Species']] == 'versicolor')
iris[['Is.Virginica']] <- as.numeric(iris[['Species']] == 'virginica')
fit.1 <- glm(Is.Versicolor ~ Petal.Length + Sepal.Length, data = iris)
fit.2 <- glm(Is.Virginica ~ Petal.Length + Sepal.Length, data = iris)
#summary(fit)
output <- capture.output(stargazer(fit.1, fit.2, title = 'Regression Results', summary=FALSE, header=FA
```


```{r echo = FALSE, results = 'asis'}
cat(output)
```

References
```

Rmarkdown basics

The following basic commands need to be executed in between the lines

```
here the Rmarkdown command lines
```

Emphasis

```
*italic*  or _italic_
```

results in

italic or *italic*

```
**bold** or  __bold__
```

results in

bold or **bold**

Headers

```
# 1st Level Header

## 2nd Level Header

### 3rd Level Header
```

results in

1st Level Header

2nd Level Header

3rd Level Header

Unordered List:

```
*   Bulleted list item 1

*   Item 2

    * Item 2a

    * Item 2b
```

results in

- Bulleted list item 1
- Item 2
  - Item 2a
  - Item 2b

Ordered List

```
1.  Numbered list item 1

1.  Item 2. The numbers are incremented automatically in the output.
```

results in

1. Numbered list item 1
2. Item 2. The numbers are incremented automatically in the output.

Links

```
<http://example.com>

[linked phrase](http://example.com)
```

results in

http://example.com

linked phrase

Images

```
![optional caption text](path/to/img.png)
```

includes the image with the optional caption text.

Tables

```
First Header  | Second Header
------------- | -------------
Content Cell  | Content Cell
Content Cell  | Content Cell
```

results in

First Header	Second Header
Content Cell	Content Cell
Content Cell	Content Cell

knitr R chunks

R code chunks can be used as a means to render R output into documents or to simply display code for illustration. Here is a simple R code chunk that will result in both the code and it's output being included:

```
```{r}
summary(cars)
```
```

This results in:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

To display the output of a code chunk but not the underlying R code, you specify the `echo = FALSE` option:

```
```{r, echo = FALSE}
summary(cars)
```
```

Note that R code chunks can also be used to render plots. To display a plot while omitting the code used to generate the plot you'd do this:

```
```{r, echo = FALSE}
plot(cars)
```
```

To display R code without evaluating it, you specify the `eval = FALSE` chunk option:

```
``{r, eval = FALSE}
summary(cars)
```
```

Table Output
```

By default data frames and matrixes are output as they would be in the R terminal (in a monospaced font). However, if you prefer that data be displayed with additional formatting you can use the knitr::kable function. For example:

```
```{r, results = 'asis'}
knitr::kable(mtcars)
```
```

results in

```
knitr::kable(mtcars)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

Note the use of the `results = 'asis'` chunk option. This is required to ensure that the raw table output isn't processed furthur by knitr. The kable function includes several options to control the maximum number of digits for numeric columns, alignment, etc (refer to the knitr package documentation for additional details).

Caching

If document rendering becomes time consuming due to long computations or plots that are expensive to generate you can use knitr caching to improve performance. The documentation knitr chunk and package options describe how caching works and the cache examples provide additional details.

If you want to enable caching globally for a document you can include a code chunk like this at the top of the document:

````
```{r setup, include = FALSE}
knitr::opts_chunk$set(cache = TRUE)
```
````

If you run into problems with cached output you can always clear the knitr cache by removing the folder named with a _cache suffix within your document's directory.

The complete list of chunk options can be found at [https://yihui.name/knitr/options/], the developer's site.

YAML Header

You can control many other "whole document" settings by tweaking the parameters of the YAML header. You might wonder what YAML stands for: it's "yet another markup language", which is designed for representing hierarchical data in a way that's easy for humans to read and write. R Markdown uses it to control many details of the output. Here we'll discuss two: document parameters and bibliographies.

The most important setting is the `output`. This defines the type of output and can be anything from a `html_document` to `beamer_presentation`

An overview of the possible settings can be found on the Rmarkdown reference card.

R Markdown Reference Guide

Learn more about R Markdown at rmarkdown.rstudio.com
Learn more about Interactive Docs at shiny.rstudio.com/articles

Contents:
1. Markdown Syntax
2. Knitr chunk options
3. **Pandoc options**

Templates

html_document
pdf_document
word_document
md_document
ioslides_presentation
slidy_presentation
beamer_presentation

Basic YAML

```
---
title: "A Web Doc"
author: "John Doe"
date: "May 1, 2015"
output: md_document
---
```

Template options

```
---
title: "Chapters"
output:
  html_document:
    toc: true
    toc_depth: 2
---
```

Latex options

```
---
title: "My PDF"
output: pdf_document
fontsize: 11pt
geometry: margin=1in
---
```

Interactive Docs

```
---
title: "Slides"
output:
  slidy_presentation:
    incremental: true
runtime: shiny
---
```

Syntax for slide formats (ioslides, slidy, beamer)

```
# Dividing slides 1

Pandoc will start a new slide at each first level header

## Header 2

… as well as each second level header

***

You can start a new slide with a horizontal rule `***` if you do not want a header.

## Bullets

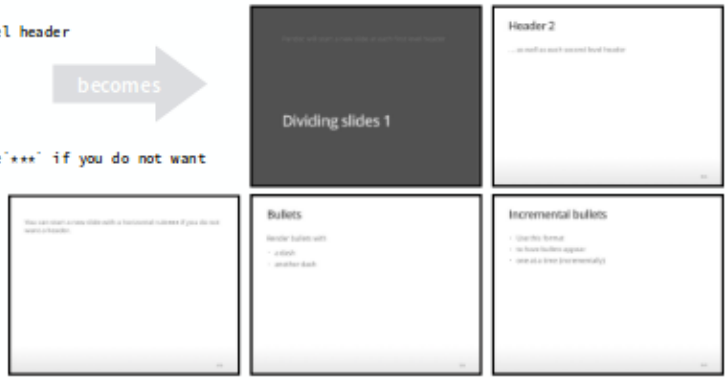
Render bullets with

- a dash
- another dash

## Incremental bullets

>- Use this format
>- to have bullets appear
>- one at a time (incrementally)
```

becomes



Slide display modes

*Press a key below during presentation to enter display mode. Press **esc** to exit display mode.*

ioslides

- **f** - enable fullscreen mode
- **w** - toggle widescreen mode
- **o** - enable overview mode
- **h** - enable code highlight mode
- **p** - show presenter notes

slidy

- **C** - show table of contents
- **F** - toggle display of the footer
- **A** - toggle display of current vs all slides
- **S** - make fonts smaller
- **B** - make fonts bigger

Top level options to customize LaTex (pdf) output

option	description
lang	Document language code
fontsize	Font size (e.g. 10pt, 11pt, 12 pt)
documentclass	Latex document class (e.g. article)
classoption	Option for document class (e.g. oneside); may be repeated
geometry	Options for geometry class (e.g. margin=1in); may be repeated
mainfont, sansfont, monofont, mathfont	Document fonts (works only with xelatex and lualatex, see the latex_engine option)
linkcolor, urlcolor, citecolor	Color for internal, external, and citation links (red, green, magenta, cyan, blue, black)

R Markdown Reference Guide

Learn more about R Markdown at rmarkdown.rstudio.com
Learn more about Interactive Docs at shiny.rstudio.com/articles

option	html	pdf	word	md	ioslides	slidy	beamer	description
colortheme							X	Beamer color theme to use (e.g., `colortheme: "dolphin"`).
css	X				X	X		Filepath to CSS style to use to style document (e.g., `css: styles.css`).
duration						X		Add a countdown timer (in minutes) to footer of slides (e.g., `duration: 45`).
fig_caption	X	X	X		X	X	X	Should figures be rendered with captions?
fig_crop		X					X	Should pdfcrop utility be automatically applied to figures (when available)?
fig_height	X	X	X	X	X	X	X	Default figure height (in inches) for document.
fig_retina	X			X	X	X		Scaling to perform for retina displays (e.g., `fig_retina: 2`).
fig_width	X	X	X	X	X	X	X	Default figure width (in inches) for document.
font_adjustmen						X		Increase or decrease font size for entire presentation (e.g., `font_adjustment: -1`).
fonttheme							X	Beamer font theme to use (e.g., `fonttheme: "structurebold"`).
footer						X		Text to add to footer of each slide (e.g., `footer: "Copyright (c) 2014 RStudio"`).
highlight	X	X				X	X	Syntax highlighting style (e.g. `"tango"`, `"pygments"`, `"kate"`, `"zenburn"`, and
includes	X	X		X	X	X	X	See below
-in_header	X	X			X	X	X	File of content to place in document header (e.g., `in_header: header.html`).
-before_body	X	X			X	X	X	File of content to place before document body (e.g., `before_body:`
-after_body	X	X			X	X	X	File of content to place after document body (e.g., `after_body: doc_suffix.html`).
incremental					X	X	X	Should bullets appear one at a time (on presenter mouse clicks)?
keep_md	X				X	X		Save a copy of .md file that contains knitr output (in addition to the .Rmd and HTML files)?
keep_tex		X					X	Save a copy of .tex file that contains knitr output (in addition to the .Rmd and PDF files)?
latex_engine		X						Engine to render latex. Should be one of `"pdflatex"`, `"xelatex"`, and `"lualatex"`.
lib_dir	X				X	X		Directory of dependency files to use (Bootstrap, MathJax, etc.) (e.g., `lib_dir: libs`).
logo					X			File path to a logo (at least 128 x 128) to add to presentation (e.g., `logo: logo.png`).
mathjax	X				X	X		Set to `local` or a URL to use a local/URL version of MathJax to render equations
number_section	X	X						Add section numbering to headers (e.g., `number_sections: true`).
pandoc_args	X	X	X	X	X	X	X	Arguments to pass to Pandoc (e.g., `pandoc_args: ["--title-prefix", "Foo"]`).
preserve_yaml				X				Preserve YAML front matter in final document?
reference_docx			X					A .docx file whose styles should be copied to use (e.g., `reference_docx:`
self_contained	X				X	X		Embed dependencies into the doc? Set to `false` to keep dependencies in external files.
slide_level							X	The lowest heading level that defines individual slides (e.g., `slide_level: 2`).
smaller					X			Use the smaller font size in the presentation?
smart	X				X	X		Convert straight quotes to curly, dashes to em-dashes, ... to ellipses, and so on?
template	X	X				X	X	Pandoc template to use when rendering file (e.g., `template:`
theme	X						X	Bootswatch or Beamer theme to use for page. Valid bootswatch themes include `"cerulean"`, `"journal"`, `"flatly"`, `"readable"`, `"spacelab"`, `"united"`, and `"cosmo"`.
toc	X	X		X			X	Add a table of contents at start of document? (e.g., `toc: true`).
toc_depth	X	X		X				The lowest level of headings to add to table of contents (e.g., `toc_depth: 2`).
transition					X			Speed of slide transitions should be `"slower"`, `"faster"` or a number in seconds.
variant				X				The flavor of markdown to use; one of `"markdown"`, `"markdown_strict"`, `"markdown_github"`, `"markdown_mmd"`, and `"markdown_phpextra"`
widescreen					X			Display presentation in widescreen format?

References

Leisch, Friedrich. 2002. "Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis." In *Compstat 2002 - Proceedings in Computational Statistics*, edited by Wolfgang Härdle and Bernd Rönz, 575–80. Physica Verlag, Heidelberg. http://www.stat.uni-muenchen.de/~leisch/Sweave.

Peng, Roger D. 2009. "Reproducible Research and Biostatistics." *Biostatistics* 10 (3): 405–8. doi:10.1093/biostatistics/kxp014.

Xie, Yihui. 2017. *Knitr: A General-Purpose Package for Dynamic Report Generation in R.* https://yihui.name/knitr/.