Writing a reproducible research paper

Julia Schulte-Cloos

2020-11-13

Abstract

Everyone agrees that this issue is really important. But we do not know much about this specific question, although it matters a great deal, for these reasons. We approach the problem from this perspective. Our research design focuses on these cases and relies on these data, which we analyse using this method. Results show what we have learned about the question. They have these broader implications.

Keywords: Bookdown, Reproducible research, Template, Manuscript

Dr. Julia Schulte-Cloos Marie Curie Research Fellow, Geschwister Scholl Institute of Political Science, Ludwig Maxmilian University of Munich, Germany. E-Mail: julia.schulte-cloos@gsi.lmu.de ORCID: 0000-0001-7223-3602

The acknowledgments go to everyone involved in creating free and open software and, in particular, to the author of the great bookdown page Yihui Xie.

The basics

This is a LaTeXbased manuscript that is generated from an Rmd-file by relying on Pandoc for conversion from Markdown to Tex. If you do not specify a `template.tex` in your `YAML` header, Pandoc will use the `default.latex` template, which you can find here. I suggest that you start writing reproducible research papers with a solid template that achieves most of your needs for producing a good-looking manuscript, but that you focus on integrating code and content over optimizing the beauty of your document. At the end of the document, you find a code chunk in which you can customize and modify the LaTeXcode that will be written to the `preamble.tex` file that we import in to our reproducible manuscript in the YAML header. I recommend that you make any changes to this preamble there to maintain a single Rmd file that you can re-use for different projects without having to drag too many single files along with the Rmd file in the different directories.

Citations

Markdown provides an easy way to cite and reference literature. We add a `bib`-file in our YAML header in the following way:

```
---
output:
  bookdown::pdf_document2:
csl: 'assets/sage-harvard.csl'
bibliography: literature.bib
link-citations: yes
---
```

We can then cite all entries included in our `.bib`-file by calling `@palmerdata.2020` for inline citations and `[@palmerdata.2020, p.10]` for all other references. Here is an example: the dataset that we use has been created by Horst et al. (2020). If our document specifies a csl style, Pandoc will convert Markdown references, i.e., `@palmerdata.2020`, to 'hardcoded' text and a hyperlink to the reference section in our document. If our document, in contrast, specifies a citation reference package like `biblatex` or `natbib` along with the related options, pandoc will create the corresponding LaTeX commands (e.g. `\autocite`, or `\pcite`) to create the references from our Markdown references.

Figures and images

There are several different ways to include images in Rmd documents. For PDF outputs, like `bookdown::pdf_document2`, we can rely on

1. Plain markdown syntax: `![A cow's nose](figs/cow.jpg){width=30%}`

Figure 1: A cow's nose

2. LaTeXsyntax

```
\begin{figure}
\centering
\includegraphics[width=0.3\textwidth]{./figs/snake.jpg}
\caption{A snake}
\end{figure}
```

Figure 2: A snake

3. Code evaluation and `knitr`

```
knitr::include_graphics(path = "figs/winter.jpg")
```

Figure 3: A winter morning

Cross-referencing

Cross-referencing is possible in PDF documents when we rely on `bookdown::pdf_document2`. We can cross-reference sections, figures, tables or equations in our document with the following syntax: `\@ref(fig:winter)`. Here is an example: Figure 3 shows a photograph of Munich

on a winter morning. If we specify the `colorlinks: true` option in our YAML header, the hyperlinks to the respective figure will be colored.

Similarly, we can also cross-reference sections, tables or equations. If you do not specify a section label, Pandoc will automatically assign a label based on the title of your header. For more details, see the Pandoc manual. If you wish to add a manual label to a header, add `{#mylabel}` to the end of the section header. If you wish to make reference to an equation, you can rely on LaTeXsyntax and put your equations in equation environments and assign a label by `(\#eq:label)`, e.g.,

```
\begin{equation}
  f\left(k\right) = \binom{n}{k} p^k\left(1-p\right)^{n-k}
  (\#eq:binom)
\end{equation}
```

Integrating code and content

Literate programming is key to reproducible documents, which means that we can integrate our code and text into a single document. We can then also include any kind of operations directly in the text by calling R with a single backtick:

```
`r (2+2)*5`
```

Here is an applied example. Let's calculate the mean bill length of penguins in the data and share this information with the readers, while rounding the number to two digits: 43.9219 mm.

Graphs - `ggplot`

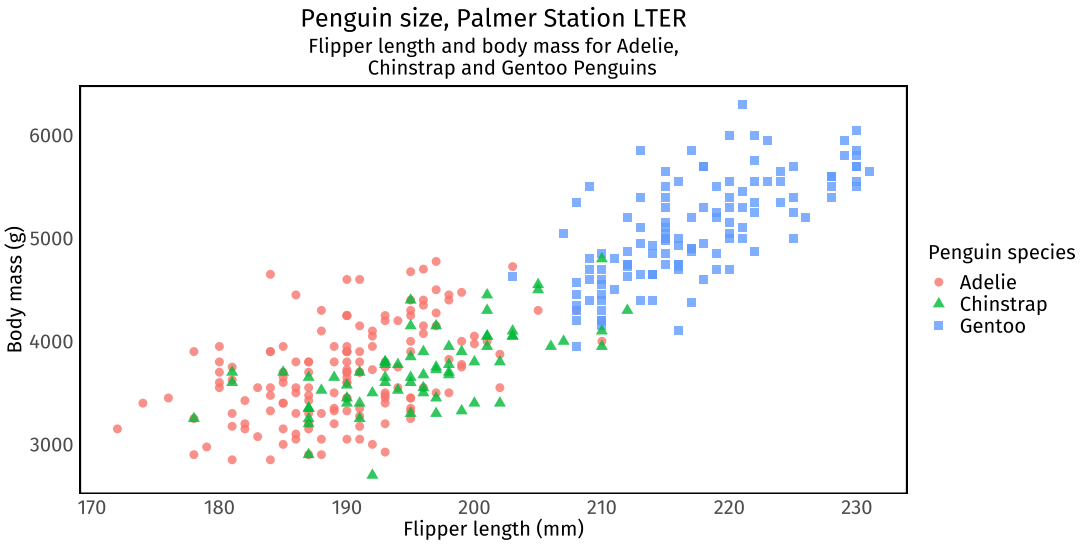

Figure 4: Body mass and flipper length of penguins.

Tables

In LaTeXdocuments, Pandoc will automatically load the packages `longtable`, `booktabs`, and `calc` when we specify the option `tables: yes` in our YAML header.

Including tables: `kable`

You can easily create and integrate your tables with the powerful table generating package `kable` and the table styling package `kableExtra`. For the full documentation of the package, see the vignette. Here is an example table including some summary statistics of the penguin species.

Table 1: Differences in Flipper and Bill Length across Penguin Species

Species	Bill Length (mm)	Bill Depth (mm)	Flipper Length (mm)	Body Mass (kg)
Adelie	38.79	18.35	189.95	3700.66
Chinstrap	48.83	18.42	195.82	3733.09
Gentoo	47.50	14.98	217.19	5076.02

Including output from regression tables: `modelsummary`

`modelsummary` is a very powerful package to present regression tables in several different output formats. Depending on your preferred table styling package, you can chose among different

output formats and then further style the regression table according to your personal needs. In the example below, we use `kableExtra` to style the table.

	Resting Pulse		Active Pulse	
	M1	M2	M3	M4
(Intercept)	86.458***	86.845***	90.769***	90.826***
	(4.507)	(4.541)	(9.397)	(9.479)
Smoke	2.048	-1.515	2.406	1.883
	(1.816)	(5.060)	(3.786)	(10.563)
Exercise	**-6.853*** **	**-7.046*** **	**-8.657*** **	**-8.685*** **
	(0.778)	(0.820)	(1.622)	(1.712)
Wgt	-0.022	-0.021	0.097*	0.097*
	(0.024)	(0.024)	(0.050)	(0.051)
Sex	1.175	1.189	9.430***	9.432***
	(1.512)	(1.513)	(3.152)	(3.159)
Smoke × Exercise		**1.915**		**0.281**
		(2.539)		(5.299)
R2	0.309	0.310	0.160	0.160
Num.Obs.	232	232	232	232

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Dataset: 'Pulse Rates and Exercise' from the Stat2Data package.

Advanced literate programming

Literate programming in figure captions

Sometimes, we would like to include the result of a specific evaluated code in the caption of a figure or a table. We can achieve this, by making use of the code chunk option `eval.after`. We might, for instance, include the overall number of penguins that are included in the dataset to our earlier Figure 4.

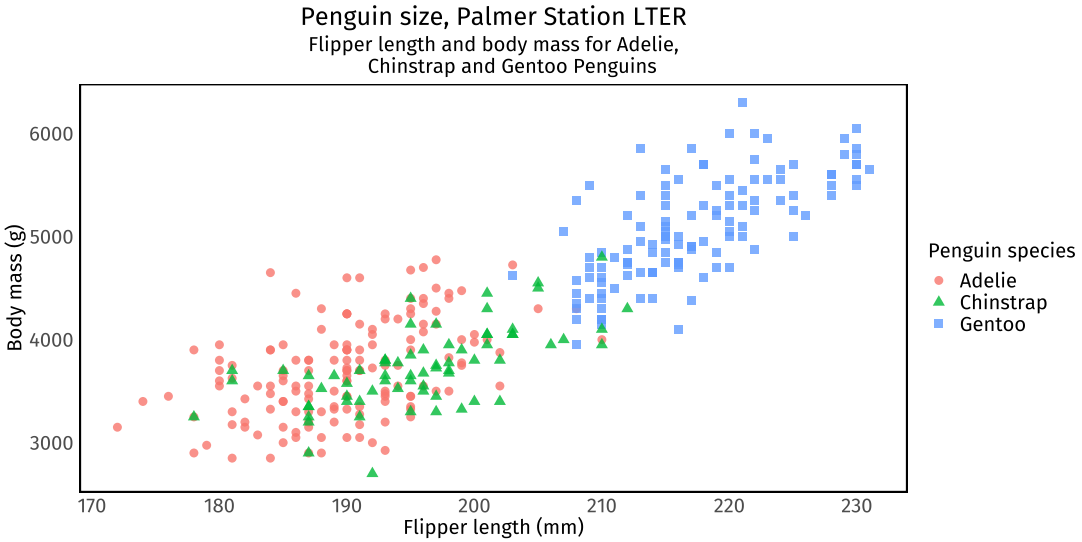

Figure 5: Body mass and flipper length of penguins. N = 344.

Working with other engines in Rmd

Python with `reticulate` package

The `reticulate` package allows two-way communication between `python` and `R`, thus, you can access any objects created or stored within a `python`-engine chunk from within an `R`-enginge chunk, and *vice versa*. To use the `python`-engine, simply replace the name of the engine after the three backticks and the curyl brace that opens a chunk.

STATA with `Statamarkdown` package

You can also use stata as an engine within your workflow. For more details, you can consult the RMarkdown cookbook (Xie et al., 2020: 15.8).

```
sysuse auto
summarize
```

Tweaks in RStudio

There are a number of useful addins in RStudio that facilitate our workflow. You should check out the `remedy` package if you would like to highlight your code, or insert chunks by point-and-click. The `styler` package is a useful addin to tidy your code, which is good practice before sharing your scripts.

References

Horst AM, Hill AP and Gorman KB (2020) *Palmerpenguins: Palmer Archipelago (antarctica) Penguin Data*. Available at: https://allisonhorst.github.io/palmerpenguins/.

Xie Y, Dervieux C and Riederer E (2020) *R Markdown Cookbook*. CRC Press.

Online appendix

Attach R session info in appendix

Since R and R packages are constantly evolving you might want to add the R session info that contains information on the R version as well as the packages that are loaded.

```
R version 4.0.3 (2020-10-10)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 18363)

Locale:
  LC_COLLATE=Portuguese_Portugal.1252  LC_CTYPE=Portuguese_Portugal.1252
  LC_MONETARY=Portuguese_Portugal.1252 LC_NUMERIC=C
  LC_TIME=Portuguese_Portugal.1252

Package version:
  askpass_1.1           assertthat_0.2.1      backports_1.1.7
  base64enc_0.1.3       BH_1.72.0.3           blob_1.2.1
  bookdown_0.21         brew_1.0.6            broom_0.7.0
  callr_3.4.4           cellranger_1.1.0      checkmate_2.0.0
  cli_2.0.2             clipr_0.7.0           colorspace_1.4-1
  commonmark_1.7        compiler_4.0.3        covr_3.5.1
  crayon_1.3.4          crosstalk_1.1.0.1     curl_4.3
  DBI_1.1.0             dbplyr_1.4.4          desc_1.2.0
  devtools_2.3.2        digest_0.6.25         dplyr_1.0.2
  DT_0.15               ellipsis_0.3.1        evaluate_0.14
  fansi_0.4.1           farver_2.0.3          forcats_0.5.0
  fs_1.5.0              generics_0.0.2        ggplot2_3.3.2
  gh_1.1.0              git2r_0.27.1          glue_1.4.2
  graphics_4.0.3        grDevices_4.0.3       grid_4.0.3
  gridExtra_2.3         gtable_0.3.0          haven_2.3.1
  highr_0.8             hms_0.5.3             htmltools_0.5.0
  htmlwidgets_1.5.1     httr_1.4.2            ini_0.3.1
  isoband_0.2.2         jsonlite_1.7.1        kableExtra_1.3.1
  knitr_1.30            labeling_0.3          later_1.1.0.1
  lattice_0.20.41       lazyeval_0.2.2        lifecycle_0.2.0
  lubridate_1.7.9       magrittr_1.5          markdown_1.1
  MASS_7.3.53           Matrix_1.2.18         memoise_1.1.0
  methods_4.0.3         mgcv_1.8.33           mime_0.9
  modelr_0.1.8          modelsummary_0.6.3    munsell_0.5.0
  nlme_3.1.149          openssl_1.4.2         pacman_0.5.1
  palmerpenguins_0.1.0 patchwork_1.1.0       pillar_1.4.6
  pkgbuild_1.1.0        pkgconfig_2.0.3       pkgload_1.1.0
  praise_1.0.0          prettyunits_1.1.1     processx_3.4.3
  progress_1.2.2        promises_1.1.1        ps_1.3.3
  purrr_0.3.4           R6_2.4.1              rcmdcheck_1.3.3
```

RColorBrewer_1.1.2	Rcpp_1.0.5	readr_1.3.1
readxl_1.3.1	rematch_1.0.1	rematch2_2.1.2
remotes_2.2.0	reprex_0.3.0	rex_1.2.0
rlang_0.4.7	rmarkdown_2.5	roxygen2_7.1.1
rprojroot_1.3-2	rstudioapi_0.11	rversions_2.0.2
rvest_0.3.5	scales_1.1.1	selectr_0.4.2
sessioninfo_1.1.1	showtext_0.9	showtextdb_3.0
splines_4.0.3	stats_4.0.3	stringi_1.5.3
stringr_1.4.0	sys_3.3	sysfonts_0.8.1
tables_0.9.6	testthat_2.3.2	tibble_3.0.3
tidyr_1.1.0	tidyselect_1.1.0	tidyverse_1.3.0
tinytex_0.26	tools_4.0.3	usethis_1.6.3
utf8_1.1.4	utils_4.0.3	vctrs_0.3.2
viridisLite_0.3.0	webshot_0.5.2	whisker_0.4
withr_2.2.0	xfun_0.19	xml2_1.3.2
xopen_1.0.0	yaml_2.2.1	