papeR

A Toolbox for Writing Pretty Papers and Reports

Benjamin Hofner
benjamin.hofner@fau.de

Department of Medical Informatics, Biometry and Epidemiology
Friedrich-Alexander-Universität Erlangen-Nürnberg

DAGStat 2016
17.03.2016
Göttingen

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Reproducible Research

An adventure with many merits (and some pitfalls)

- Repeated observations, repeated measurements, and replications of experiments and studies are *essential for modern science*.

> *"Science moves forward when discoveries are replicated and reproduced."*
>
> Stodden et al. [2014]

- Reproducible research (*RR*): publications should be accompanied by all relevant material to reproduce the results and findings
- Claerbout's principle: Articles are merely the advertisement of the underlying code and data.

- Thus, RR boosts our scientific reputation (at least it should).
- Yet, it is (sometimes) difficult to achieve.

Reproducible research in statistics: A review and guidelines for the *Biometrical Journal*

Benjamin Hofner[*,1], **Matthias Schmid**[2], and **Lutz Edler**[3]

[1] Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander University, Erlangen-Nuremberg, Waldstraße 6, 91054 Erlangen, Germany
[2] Department of Medical Biometry, Informatics and Epidemiology, University of Bonn, Sigmund-Freud-Straße 25, 53127 Bonn, Germany
[3] Division of Biostatistics-C060, German Cancer Research Center, Im Neuenheimer Feld 581, 69120 Heidelberg, Germany

Received 31 July 2015; revised 6 October 2015; accepted 17 November 2015

Reproducible research (RR) constitutes the idea that a publication should be accompanied by all relevant material to reproduce the results and findings of a scientific work. Hence, results can be verified and researchers are able to build upon these. Efforts of the *Biometrical Journal* over the last five years have increased the number of manuscripts which are reproducible by a factor of 4 to almost 50%. Yet, more than half of the code submission could not be executed in the initial review due to missing code, missing data or errors in the code. Careful checks of the submitted code as part of the reviewing process are essential to eliminate these issues and to foster RR. In this article, we reviewed $n = 56$ recent submissions of code and data to identify common reproducibility issues. Based on these findings, guidelines for structuring code submission to the *Biometrical Journal* have been established to help authors. These guidelines should help researchers to implement RR in general. Together with the code reviews, this supports the mission of the *Biometrical Journal* in publishing highest quality, novel and relevant papers on statistical methods and their applications in life sciences. Source code and data to reproduce the presented data analyses are available as Supplementary Material on the journal's web page.

Reproducible research in statistics: A review and guidelines for the *Biometrical Journal*

Benjamin Hofner[*,1], **Matthias Schmid**[2], and **Lutz Edler**[3]

[1] Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander University, Erlangen-Nuremberg, Waldstraße 6, 91054 Erlangen, Germany
[2] Department of Medical Biometry, Informatics and Epidemiology, University of Bonn, Sigmund-Freud-Straße 25, 53127 Bonn, Germany
[3] Division of Biostatistics-C060, German Cancer Research Center, Im Neuenheimer Feld 581, 69120 Heidelberg, Germany

Received 31 July 2015; revised 6 October 2015; accepted 17 November 2015

Reproducible research (RR) constitutes the idea that a publication should be accompanied by all relevant material to reproduce the results and findings of a scientific work. Hence, results can be verified and researchers are able to build upon these. Efforts of the *Biometrical Journal* over the last five years have increased the number of manuscripts which are reproducible by a factor of 4 to almost 50%. Yet, more than half of the code submission could not be executed in the initial review due to missing code, missing data or errors in the code. Careful checks of the submitted code as part of the reviewing process are essential to eliminate these issues and to foster RR. In this article, we reviewed $n = 56$ recent submissions of code and data to identify common reproducibility issues. Based on these findings, guidelines for structuring code submission to the *Biometrical Journal* have been established to help authors. These guidelines should help researchers to implement RR in general. Together with the code reviews, this supports the mission of the *Biometrical Journal* in publishing highest quality, novel and relevant papers on statistical methods and their applications in life sciences. Source code and data to reproduce the presented data analyses are available as Supplementary Material on the journal's web page.

One solution: "Literate Programming"

- The name *Literate Programming* (mixing the source code with documentation) was first coined by Knuth [1984].
- Instead of copy and paste output, figures, etc. to your document, they are part of the document.
  - It is by far less error prone.
  - It is easier to update the report, e.g., if the data changes.
  - In summary it is often faster to use literate programming, even if it requires more work in the first place.
- In R there exists a variety of tools such as `Sweave` [Leisch, 2002], or **knitr** [Xie, 2014, 2015].

More Details

- see e.g. `http://www.benjaminhofner.de/downloads/2015/RR_short_course/Using_knitr.pdf`

Statistical Reporting

- Most of the work for (and space in) a statistical report is needed for standardized reporting.
- Repeated tasks in a statistical report:
  - Summary tables
  - Grouped summary tables
  - Simple graphical displays
  - Display of model results
- To ease statistical reporting (especially in in the framework of literate programming)
  - Use the R package **papeR** [Hofner, 2015]
- This frees statisticians to think about the important issues of the statistical analysis.

Package papeR

The package ...

- uses variable labels if provided, e.g., after import from SPSS (`labels()`)
- easily creates summary tables (`summarize()`)
- provides a special plot function for labeled data frames (`plot()`)
- allows to enhance and prettify summary tables of statistical models by (possibly) adding
  - confidence intervals,
  - significance stars,
  - odds ratios, etc.

  and by separating variable names and factor levels (`prettify()`).

Setting and Extracting Labels

```r
## Load package
library("papeR")
## Load a data example
data(Orthodont, package = "nlme")
## Per default label = variable name:
labels(Orthodont)

##   distance        age    Subject        Sex
## "distance"      "age"  "Subject"      "Sex"

## Set variable labels
labels(Orthodont) <- c("Fissure distance (mm)",
                       "Age (years)", "Subject", "Sex")
## Show labels
labels(Orthodont)

##                  distance                       age
## "Fissure distance (mm)"           "Age (years)"
##                   Subject                       Sex
##                 "Subject"                     "Sex"
```

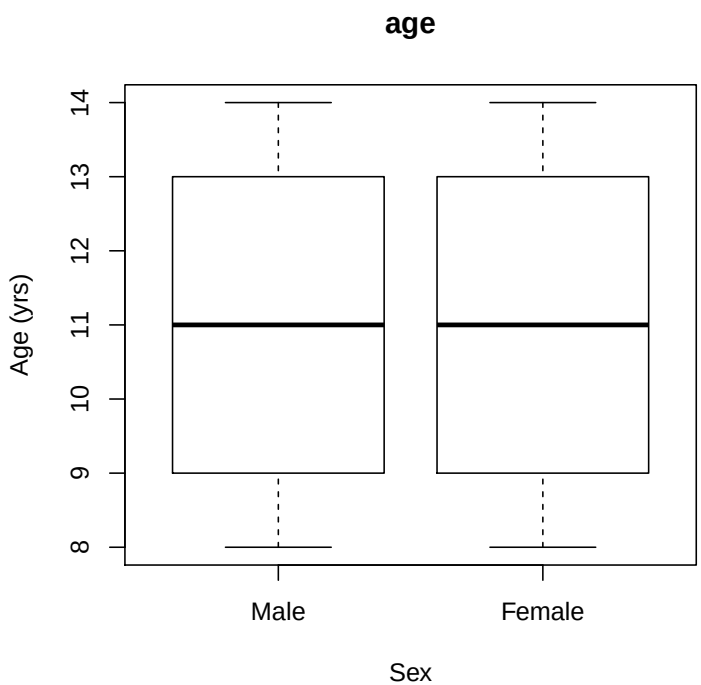
Advanced Labeling

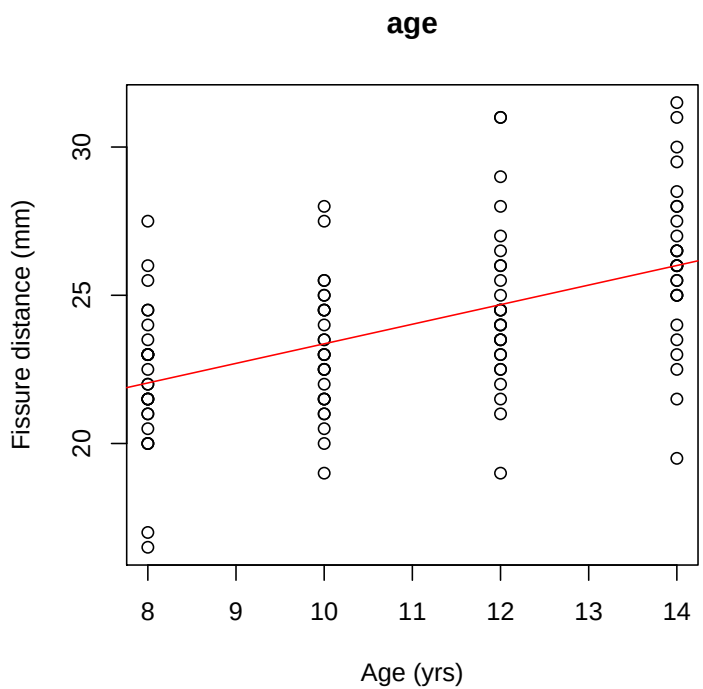
```
## Show label for age
labels(Orthodont, which = 2)

##           age
## "Age (years)"

## The same
labels(Orthodont, which = "age")

##           age
## "Age (years)"

## One can also set the label for age
labels(Orthodont, which = "age") <- "Age (yrs)"
```

Plotting

```
## Standard plot for each variable (not shown here)
plot(Orthodont)
```

```
## Bivariate plots
par(mfrow = c(1, 2))
plot(Orthodont, variables = "age", by = "Sex")
plot(Orthodont, variables = "age", with = "distance")
```





Summary statistics

```
summarize(Orthodont, type = "numeric")
```

```
##              N    Mean   SD    Min Q1 Median Q3  Max
## 1 distance 108   24.02 2.93   16.5 22  23.75 26 31.5
## 2      age 108   11.00 2.25    8.0  9  11.00 13 14.0
```

```
summarize(Orthodont, type = "factor", variables = "Sex")
```

```
##        Level    N    %
## 1 Sex   Male   64 59.3
## 2     Female   44 40.7
```

Grouped summary statistics

```
summarize(Orthodont, type = "numeric", group = "Sex")
```

```
##                Sex    N    Mean   SD    Min Q1 Median    Q3  Max   p.value
## 1 distance    Male   64   24.97 2.90   17.0 23  24.75 26.50 31.5    <0.001
## 2          Female   44   22.65 2.40   16.5 21  22.75 24.25 28.0
## 3      age    Male   64   11.00 2.25    8.0  9  11.00 13.00 14.0     1.000
## 4          Female   44   11.00 2.26    8.0  9  11.00 13.00 14.0
```

- Per default a t-test is computed.
- Different tests can be used as well.
- Grouped statistics also exist for `type = "factor"` (with Fisher test per default).

PDF output for Summaries

- `summarize` produces an informative console output.
- Yet, for reports one wants a proper table.

PDF output for Summaries

- `summarize` produces an informative console output.
- Yet, for reports one wants a proper table.
- Simply use `xtable`:

```
library("xtable")
xtable(summarize(Orthodont, type = "numeric"))
```

	N	Mean	SD	Min	Q1	Median	Q3	Max
distance	108	24.02	2.93	16.50	22.00	23.75	26.00	31.50
age	108	11.00	2.25	8.00	9.00	11.00	13.00	14.00

```
xtable(summarize(Orthodont, type = "factor", variables = "Sex"))
```

	Level	N	%
Sex	Male	64	59.3
	Female	44	40.7

```
xtable(summarize(Orthodont, type = "numeric",
                 quantiles = FALSE, group = "Sex"))
```

	Sex	N	Mean	SD	p.value
distance	Male	64	24.97	2.90	<0.001
	Female	44	22.65	2.40	
age	Male	64	11.00	2.25	1.000
	Female	44	11.00	2.26	

- The tables can be also converted to Markdown, e.g. using `knitr::kable()`.

Prettify model output

```
linmod <- lm(distance ~ age + Sex, data = Orthodont)
## Extract pretty summary
(pretty_lm <- prettify(summary(linmod)))

##                 Estimate CI (lower) CI (upper) Std. Error
## 1 (Intercept) 17.7067130 15.5014071 19.9120189 1.11220946
## 2         age  0.6601852  0.4663472  0.8540231 0.09775895
## 3 Sex: Female -2.3210227 -3.2031499 -1.4388955 0.44488623
##     t value Pr(>|t|)
## 1 15.920304   <0.001 ***
## 2  6.753194   <0.001 ***
## 3 -5.217115   <0.001 ***
```

- As can be seen, CIs are added,
- p-values are properly formated,
- and significance stars are added.
- Furthermore, variable name and label are separated.
- Again, `xtable` or `kable` can be used for proper tables.

Pretty models

Compare standard summary table

```
xtable(summary(linmod))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.7067	1.1122	15.92	0.0000
age	0.6602	0.0978	6.75	0.0000
SexFemale	-2.3210	0.4449	-5.22	0.0000

to pretty table

```
xtable(prettify(summary(linmod)))
```

	Estimate	CI (lower)	CI (upper)	Std. Error	t value	Pr(>|t|)	
(Intercept)	17.71	15.50	19.91	1.11	15.92	<0.001	***
age	0.66	0.47	0.85	0.10	6.75	<0.001	***
Sex: Female	-2.32	-3.20	-1.44	0.44	-5.22	<0.001	***

Summary & Outlook

- Reproducible research is a basic foundation of science.
- Literate programming eases reproducibility.
- **papeR** eases statistical reporting.
- Results can be used in LaTeX and Markdown reports.

Further resources:

- Package and tutorials with more examples are available as vignettes on CRAN:
  `https://cran.r-project.org/package=papeR`
- Feature requests, bug reports and latest package version:
  `http://github.com/hofnerb/papeR`

References I

Benjamin Hofner. *papeR: A Toolbox for Writing Pretty Papers and Reports*, 2015. URL `http://CRAN.R-project.org/package=papeR`. R package version 1.0-0.

Donald Ervin Knuth. Literate programming. *The Computer Journal*, 27 (2):97–111, 1984.

Friedrich Leisch. Sweave, Part I: Mixing R and LaTeX. *R News*, 2(3): 28–31, 2002.

Victoria Stodden, Friedrich Leisch, and Roger D Peng. *Implementing reproducible research*. CRC Press, 2014.

Yihui Xie. *Dynamic Documents with R and knitr*. Chapman & Hall, CRC Press, 2014.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2015. URL `http://CRAN.R-project.org/package=knitr`. R package version 1.9.