

Dynamic reporting with R/Sweave and L^AT_EX

Uwe Ziegenhagen

Abstract

R is a sophisticated statistical programming language available for various platforms. Since its initial development in 1992 it has become the major open source tool for statistical data analysis. For the integration with L^AT_EX it provides tools which allow the convenient dynamic creation of reports. In this article I will give a very brief introduction to R and show how R integrates in the L^AT_EX workflow.

1 Introduction to R

The history of R dates back to 1969 when John M. Chambers from Bell Labs published the outlines of S, a programming language for statistics and data analysis that was first implemented in 1975 for Honeywell computers. Starting in 1992 Ross Ihaka and Robert Gentleman from the University of Auckland in New Zealand took up the concepts underlying S to develop R, a free implementation of the language. Today R is for many statisticians the tool of choice for visualization and data analysis, covering all aspects of modern computer-based statistics. The R project team has more than 500 members; more than 1,000 packages are available on CRAN, the Comprehensive R Archive Network.

1.1 R as a calculator

Since the focus of this paper lies more on the interaction with L^AT_EX this is not the place for a thorough introduction to R. Interested readers may have a closer look in the bibliography of this article for suitable materials. Nevertheless I would like to point out some of the main features of the language. Listing 1.1 shows some of the operators for basic calculations.

```

1 1+2
2 1*2
3 1/2
4 1-2
5 2^2
6 sqrt(2)
7 sin(pi) # cos, tan
8 trunc(-pi) # -3
9 round(pi) # 3

```

Listing 1.1: Basic calculations with R

The basic objects to store variables in R are vectors, matrices and dataframes. Vectors and matrices may contain only a single data type; complex data struc-

tures are stored in so-called dataframes, which are in fact lists of objects that may have different data types. Various ways of creating and assigning vectors to variables are shown in Listing 1.2.

```

1 a <- 1:3 # store vector 1..3 in a
2 b = 2:4 # store 2..4 in b
3 c(a,b) # [1] 1 2 3 2 3 4 # cat a & b
4 # generate sequence
5 seq(1,2,by=0.1) [1] 1.1 1.2 1.3 ...
6 # repeat 1..4 twice
7 rep(1:4,2) # [1] 1 2 3 4 1 2 3 4

```

Listing 1.2: Generating vectors in R

Our final R example in Listing 1.3 shows how a simple linear model can be computed with R. The vector of independent variables x contains the numbers 1 to 10; for the vector of dependent variables y we just multiply the x -vector with a random factor taken from a normal distribution. The linear model is then calculated using the `lm` command which presents the coefficients of the linear model.

```

1 > x<-1:10
2 > y=rnorm(10)*x
3 > lm(y~x)
4
5 Call:
6 lm(formula = y ~ x)
7
8 Coefficients:
9 (Intercept)          x
10 0.1079          1.0697

```

Listing 1.3: A linear model with R

1.2 R graphics

R manages its graphical output (see the basic example code in Listing 1.4 and its output in Figure 1) through *graphics devices* which take the graphics object data and convert them into printable or viewable form. The list of available graphics devices is extensive; there are devices for PDF, PostScript, X11, Java and SVG, to mention a few. Listing 1.5 shows for example how the PDF device can be used to produce high-quality PDF files.

```

1 a<- c(1:10)
2 plot(a)

```

Listing 1.4: The most basic R plot

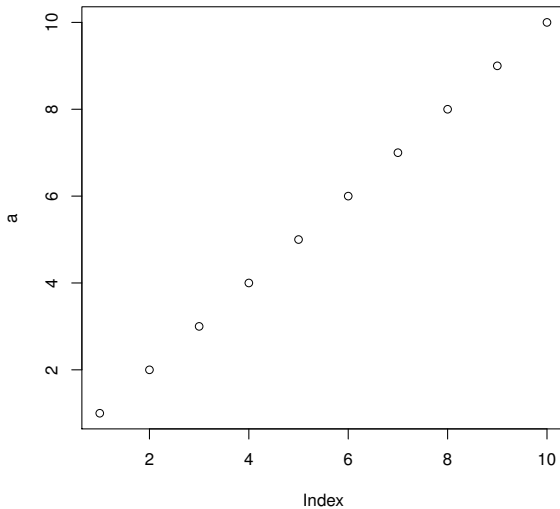


Figure 1: Graphics generated by Listing 1.4

```

1 pdf(file = "c:/punkte.pdf",width = 6,
2   height = 6, onefile = FALSE,
3   family = "Helvetica",
4   title = "R Graphics Output",
5   fonts = NULL, version = "1.4",
6   paper = "special")
7
8 a<- c(1:10)
9 plot(a)
10
11 # switch back to screen device
12 dev.off()

```

Listing 1.5: Example code for the PDF device

1.3 The TikZ graphics device

Especially interesting for T_EXnicians is the TikZ device which generates source code for the L^AT_EX graphics package of the same name. This device either creates files that can be compiled standalone or just the graphics code to be embedded in a L^AT_EX document. The advantage of this device, compared with others, is that the internal fonts of the document are used and mathematical code may be used in captions and labels as well. Listing 1.7 shows an excerpt from the file generated by the code from Listing 1.6.

```

1 tikz(file = "c:/test2.tex",standAlone=F)
2 # StandAlone=T
3 plot(1:10)
4
5 dev.off()

```

Listing 1.6: Example code for the TikZ device

```

1 % Created by tikzDevice
2 \begin{tikzpicture}[x=1pt,y=1pt]
3 \draw[color=white,opacity=0] (0,0)
4   rectangle (505.89,505.89);
5 \begin{scope}
6 \path[clip] ( 49.20, 61.20) rectangle (480.69,456.69);
7 \definecolor[named]{drawColor}{rgb}{0.56,0.96,0.51}
8 \definecolor[named]{fillColor}{rgb}{0.13,0.09,0.52}
9 \definecolor[named]{drawColor}{rgb}{0.00,0.00,0.00}
10 \draw[color=drawColor,line cap=round,line join=round,
11 fill opacity=0.00,] ( 65.18, 75.85) circle ( 2.25);
12 \draw[color=drawColor,line cap=round,line join=round,
13 fill opacity=0.00,] (109.57,116.54) circle ( 2.25);
14 \end{scope}
15 \begin{scope}

```

Listing 1.7: Excerpt from the generated TikZ code

2 Sweave and R

2.1 Introduction

In the second part of the article I want to introduce the Sweave package, developed by Friedrich Leisch. Sweave is part of the standard R installation so it requires no additional effort to install.

The package allows to include both R and L^AT_EX code in a single file. The R code is enclosed in “noweb” tags, <<>>= for the beginning, @ for the end. Noweb is a free tool implementing Donald Knuth’s approach of literate programming (for more details on this topic please see the articles in Wikipedia, et al.). The noweb-file is then processed within R using the command `Sweave("<filename>")`. To extract the R code from the file, Sweave also provides a second command, `Stangle`.

Listing 2.1 shows a very basic example, calculating 1 + 1.

```

1 \documentclass{article}
2 \begin{document}
3 <<>>=
4 1+1
5 @
6 \end{document}

```

Listing 2.1: Basic Sweave example

When we process the file from Listing 2.1 using the `Sweave` command in R we receive the L^AT_EX document shown in Listing 2.2. This document can then be compiled to PDF or DVI shown in Figure 2. As we can see in the document, Sweave requires the L^AT_EX package of the same name which provides commands for the input and output of Sweave code to be in the search path.

2.2 Sweave options

Sweave allows various options to be set within the <<>>= tag; for example, `echo=false` suppresses the

```

1 \documentclass{article}
2 \usepackage{Sweave}
3 \begin{document}
4 \begin{Schunk}
5 \begin{Sinput}
6 > 1 + 1
7 \end{Sinput}
8 \begin{Soutput}
9 [1] 2
10 \end{Soutput}
11 \end{Schunk}
12 \end{document}

```

Listing 2.2: L^AT_EX generated by Listing 2.1

Figure 2: PDF file generated by Listing 2.1

output of the original R source, while `results=hide` suppresses the output of results. A combination of both options may not seem to make sense, but this can be used to load data in the beginning of the analysis, set default values for variables, etc.

Since there are R packages that directly create valid L^AT_EX source, Sweave supports `results=tex`, a mode that passes the generated output through to L^AT_EX without tampering with its content.

If images are created in an R code chunk the option `fig=true` needs to be set. The default setting is to create both PostScript and PDF versions of each plot; the settings `eps=true/false` and `pdf=true/false` adjust this, respectively. Finally, the size of the plot can be set using the `width` and `height` parameter specifying the width and height of each plot in inches. Options may also be set globally by `\SweaveOpts`; see the Sweave manual for details.

Sweave also implements the noweb way of re-using code chunks: a certain piece of code can be named with `<<<name>>`, `opt=...>>=`; the user may then use these chunks with `<<<name>>>`.

For scalar results which can be, for example, embedded in the running text, Sweave provides the `\Sexpr(<R-code>)` command. The only requirement for the code is that the return value must be either a string or an object that can be converted to a string. We will use this command in the example described next, shown in Listing 2.3.

```

1 \documentclass[a4paper]{scrartcl}
2 \begin{document}
3 <<echo=false,results=hide>>=
4 data(iris) # load iris data
5 @
6 The data set has \Sexpr{ncol(iris)} columns
7 and \Sexpr{nrow(iris)} rows.
8
9 <<echo=false>>=summary(iris$Petal.Length)
10 @
11 <<echo=false,results=tex>>=
12 xtable(lm(iris$Sepal.Width~iris$Petal.Length),
13 caption="Linear Model of Sepal.Width
14 and Petal.Length")
15 @
16 \centering
17 \begin{figure}[htp]
18 <<fig=true,echo=false>>=
19 pch.vec <- c(16,2,3)[iris$Species]
20 col.vec <- c(16,2,3)[iris$Species]
21 plot(iris$Sepal.Width,iris$Petal.Length,
22 col = col.vec,pch=pch.vec)
23 @
24 \caption{Plot of iris\$Petal.Length vs. iris
25 \\\$Sepal.Width}
26 \end{figure}
27 \end{document}

```

Listing 2.3: Sweave code to generate Figure 3

2.3 The iris example

Listing 2.3 shows a brief example for a statistical analysis of the well-known iris dataset, consisting of each 50 observations for three different species of iris flowers. In the first R code chunk the data is loaded, since this step may not be relevant for the reader we omit both the output and the R code.

To print the number of rows and columns for this dataset we use the `\Sexpr()` command in the L^AT_EX text before displaying a small summary of the data. Afterwards we have R compute the linear model for two variables and print the results using R's `xtable` command, which provides output in valid L^AT_EX syntax. Therefore we prevent Sweave from displaying the code in verbatim mode by specifying `results=tex`. Finally, the last code chunk plots a scatterplot for the variables `Petal.length` and `Sepal.width`; please note the `fig=true` statement here, specifying that the result is a picture.

2.4 Dynamic reports

The final example shows how reports with dynamic data sources can be created easily. Let's suppose we need a report on the USD/EURO exchange rates on a frequent basis. The data can be downloaded

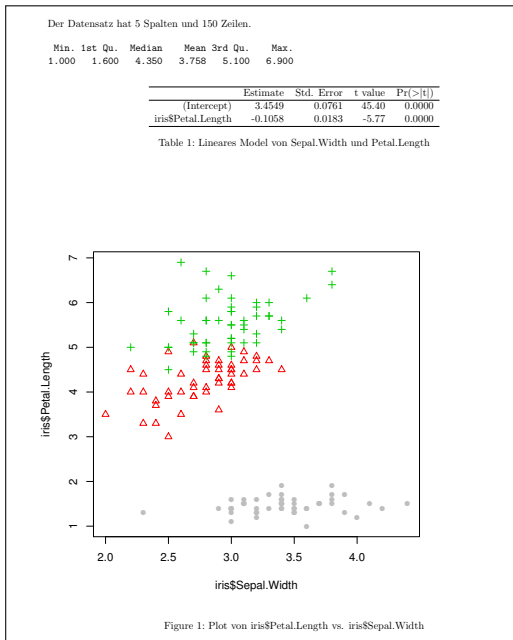


Figure 3: Document generated by Listing 2.3

from the home page of the European Central Bank where it is provided in XML or CSV format. The retrieval process can be controlled from R using the `system()` command calling an external `wget` (standard on Linux/Unix; Windows users may need to install it). To extract the data from the resulting zip file R's internal zip tool is used and the data set stored in the variable `data`. The code is shown in Listing 2.4.

We use the `\Sexpr()` command to print the dimensions of the dataset in our document before plotting a chart (see Figure 4) with the development of the Euro/dollar exchange rate.

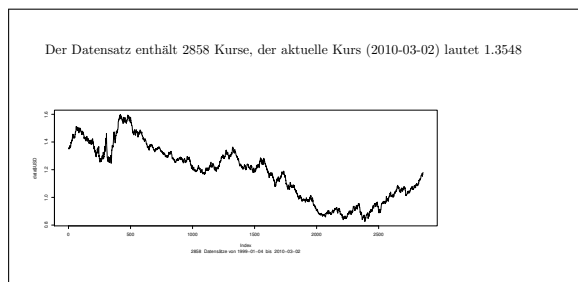


Figure 4: Output of exchange rate example, generated by Listing 2.4

3 Conclusion

R provides a vast set of functions for all aspects of data analysis. Together with Sweave, a user can easily generate dynamic reports holding the results and methods leading to them in one document. With

Uwe Ziegenhagen

```

1 \documentclass{scrartcl}
2 \begin{document}
3 <<echo=f,results=hide>>=
4 # define width and height of the plot window
5 windows(width = 8, height = 4)
6 # use external wget to download the file from
7 # the ECV and to store it in d.zip
8 system("wget -O d.zip http://www.ecb.int/
9 stats/eurofxref/eurofxref-hist.zip")
10 # use R's internal zip to extract the file to
11 # the current directory
12 zip.file.extract(file="eurofxref-hist.csv",
13 zip="d.zip", unzip="", dir=getwd())
14 # read the data
15 data= read.csv("eurofxref-hist.csv", sep=",",
16 header=TRUE)
17 @
18 The data contains \Sexpr{nrow(data)} rates,
19 the latest rate (\Sexpr{data$Date[1]}) was
20 \Sexpr{data$USD[1]}
21 \centering
22 \begin{figure}[ht]
23 <<fig=true,echo=false,width=15,height=6>>=
24 # plot the data
25 plot(data$USD,t="1", sub=paste(nrow(data),"
26 datasets from ",data$Date[nrow(data)],"
27 until ",data$Date[1]),asp=)
28 @
29 \end{figure}
30 \end{document}

```

Listing 2.4: Sweave code with dynamic data source

this article I want to encourage everybody to give R a try when facing a data analysis challenge. If any questions or remarks, please feel free to contact me.

References

- [1] Michael J. Crawley, *Statistics — An Introduction using R*. Wiley, 2005.
- [2] Peter Dalgaard, *Introductory Statistics with R*. Springer-Verlag, 2004.
- [3] Uwe Ligges, *Programmieren mit R*. Springer-Verlag, 2008.
- [4] John Maindonald and John Brown, *Data Analysis and Graphics Using R*. Cambridge University Press, 2010.
- [5] R Core Team, *An Introduction to R*. <http://cran.r-project.org/doc/manuals/R-intro.pdf>

◇ Uwe Ziegenhagen
Lokomotivstr. 9
50733 Cologne, Germany
ziegenhagen (at) gmail dot com
<http://www.uweziegenhagen.de>