Dynamic documents with R and LaTeX as an important part of reproducible research

Bernd Weiss
Research Institute for Sociology
University of Cologne
Germany

02/11/2011

Outline

Preliminary remarks

Reproducible research

LATEXin 5 minutes

Sweave

References and additional materials

Acknowledgment, license and downloads

- My presentation was created using Emacs' *org-mode* and *Babel: active code in Org-mode*.

- Slides, dataset and R code can be downloaded from my github page: (see "Downloads" button on the right-hand side).

Topic

Outline

- Present some ideas about reproducible research
- My data analysis workflow
- LaTeX in 5 minutes
- Sweave: Dynamic documents with R and LaTeX

Topic

What is reproducible research?

"By reproducible research, we mean research papers with accompanying software tools that allow the reader to directly reproduce the results and employ the methods that are presented in the research paper" (Gentleman/Lang 2004: 1).

Requirements for the workflow: TREMMP

- Transparency (e.g., by using dynamic documents, "The source code is real")

- Reproducibility (e.g., by using dynamic documents, "The source code is real")

- Efficiency (a good workflow saves you time, by automating as much of the process as possible)

- Maintainability (standardized script names, good commenting practices, README files)

- Modularity (discrete tasks into separate components (e.g. scripts))

- Portability (e.g., by using relative (not absolute) pathnames)

(Source: David Smith on "A workflow for R":
http://blog.revolutionanalytics.com/2010/10/a-workflow-for-r.html)

Sweave as a tool for creating dynamic documents

"Sweave is a tool that allows the R code used for a complete data analysis to be embedded into LaTeX, HTML or OpenOffice documents. The purpose is to create dynamic reports, which can be updated automatically if the data or analysis change. Instead of inserting a prefabricated graph or table into the report, the master document contains the R code necessary to obtain it. When run through R, all data analysis output (tables, graphs, etc.) is created on the fly and inserted into a final document. Data, code and documentation are tightly linked together, which allows for truly reproducible research" (Leisch 2010, http://user2010.org/tutorials/Leisch.html).

Some Sweave examples

- Dissertation thesis
- Medical report with tables and figures

The source code is real

"The source code is real. The objects are realizations of the source code. Source for EVERY user modified object is placed in a particular directory or directories, for later editing and retrieval" (Rossini et al. 2011:ESS - Emacs Speaks Statistics - Manual)

How my project(s) are organized

```
 1  total 16
 2  -rw-r--r-- 1 1073 Nov  5 08:33 README
 3  drwxr-xr-x 1 4096 Jan 30 10:14 data/
 4  drwxr-xr-x 1    0 Jan 26 12:13 docs/
 5  drwxr-xr-x 1    0 Feb  4 08:11 graphs/
 6  drwxr-xr-x 1    0 Jan 29 16:39 libs/
 7  drwxr-xr-x 1    0 Dec 23 09:34 org/
 8  drwxr-xr-x 1 4096 Feb  4 08:11 reports/
 9  drwxr-xr-x 1 4096 Jan 30 08:03 src/
10  drwxr-xr-x 1    0 Jan 29 11:44 tests/
```

Topic

What is LaTeX

- LaTeX is a markup language. Another markup language you might know is HTML.
- LaTeX provides high-quality typesetting features.
- The typical workflow is as follows:
  1. Create LaTeX source code file (`.tex`)
  2. Compile it via LaTeX or pdfLaTeX
  3. Use a viewer (PDF, DVI or via dvips PS) to view the compiled file
- In order to run LaTeX on your computer, you will need to install a LaTeX-distribution (e.g., MikTeX for MS-Windows).

The TEX work flow
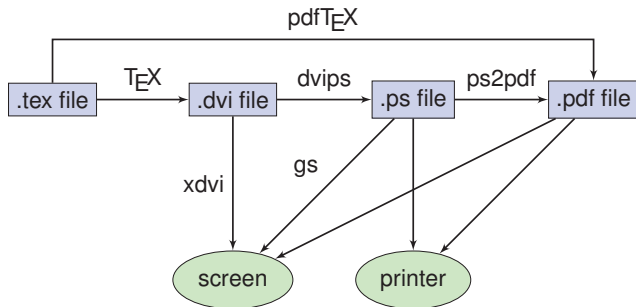
Source:
http://media.texample.net/tikz/examples/PDF/tex-
workflow.pdf

What a LaTeX file looks like

```latex
1   %% Part 1: Preamble
2   \documentclass{article}
3
4   \usepackage[utf8]{inputenc}
5   \usepackage[T1]{fontenc}
6   \usepackage[english]{babel}
7
8   %% Part 2: Body
9   \begin{document}
10
11  \section{Heading}
12
13  Hello world!
14
15  \begin{equation}
16  \overline{T} = \frac{\sum\limits^{k}_{i = 1} %
17    T_{i}\cdot w_{i}}{\sum\limits^{k}_{i = 1}w_{i}}
18  \end{equation}
19
20  \end{document}
```

The compiled 'Hello world'-example

1   Heading

Hello world!

$$\overline{T} = \frac{\sum\limits_{i=1}^{k} T_i \cdot w_i}{\sum\limits_{i=1}^{k} w_i} \qquad (1)$$

Topic

What a Sweave file looks like

```
%% filename: src/tex/examp_sweave-01.Rnw
\documentclass[noae]{article}

\usepackage[utf8x]{inputenc}
\usepackage[T1]{fontenc}
\usepackage[english]{babel}
\usepackage[margin = 1in]{geometry}

\title{This is a tiny Sweave example}
\author{Bernd Weiss}

\begin{document}
\maketitle

I am using a built-in dataset which is called \texttt{USArrests}
(Violent Crime Rates by US State).

<<echo = TRUE>>=
summary(USArrests)
@

The mean for "Murder arrests (per 100,000)" is \Sexpr{mean(USArrests$Murder)}.

\setkeys{Gin}{width=0.4\textwidth}

\begin{figure}[h!]
\begin{center}
<<echo = FALSE, fig = TRUE>>=
hist(USArrests$Murder)
@
\end{center}
\caption{Murder arrests (per 100,000)}
\end{figure}

\end{document}
```

Running Sweave

```
1  setwd("E:/projects/software/reproducible_research/src/te
2  Sweave("examp_sweave-01.Rnw")
3  system("pdflatex -output-directory ../../slides examp_sw
4          show.output.on.console = TRUE,
5          minimized = FALSE)
```

The compiled Sweave-example

This is a tiny Sweave example

Bernd Weiss

February 11, 2011

I am using a built-in dataset which is called USArrests (Violent Crime Rates by US State).

```
> summary(USArrests)
```

```
     Murder          Assault         UrbanPop          Rape
 Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
 1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
 Median : 7.250   Median :159.0   Median :66.00   Median :20.10
 Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
 3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
 Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
    Murder_d     UrbanPop_d
 (0.8,7.25]:24   (32,66]:25
 (7.25,17.4]:25  (66,91]:24
 NA's    : 1     NA's   : 1
```

The mean for "Murder arrests (per 100,000)" is 7.788.

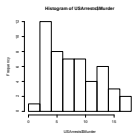**Histogram of USArrests$Murder**

Figure 1: Murder arrests (per 100,000)

1

A second Sweave example

```
1  %% filename: src/tex/examp_sweave-02.Rnw
2  \documentclass[noae]{article}
3
4  \usepackage[utf8x]{inputenc}
5  \usepackage[T1]{fontenc}
6  \usepackage[english]{babel}
7
8  \title{This is a tiny Sweave example}
9  \author{Bernd Weiss}
10
11 \begin{document}
12 \maketitle
13
14 How to create publication-ready table:
15
16 <<echo = FALSE, results = tex>>=
17 library(xtable)
18 USArrests$Murder_d <- cut(USArrests$Murder,
19                           quantile(USArrests$Murder,
20                           probs = seq(0, 1, 0.5)))
21 USArrests$UrbanPop_d <- cut(USArrests$UrbanPop,
22                             quantile(USArrests$UrbanPop,
23                             probs = seq(0, 1, 0.5)))
24 xtable(table(USArrests$Murder_d, USArrests$UrbanPop_d))
25 @
26
27
28 <<echo = FALSE, results = tex>>=
29
30 xtable(lm(Murder ~ UrbanPop, data = USarrests))
31 @
32
33 \end{document}
```

The second compiled Sweave-example

This is a tiny Sweave example

Bernd Weiss

February 11, 2011

How to create publication-ready table:

	(32,66]	(66,91]
(0.8,7.25]	13	10
(7.25,17.4]	11	14

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.4159	2.9067	2.21	0.0321
UrbanPop	0.0209	0.0433	0.48	0.6312

1

Topic

(Some) R packages that generate LaTeX code

- Hmisc: Harrell Miscellaneous
- memisc: Tools for Management of Survey Data, Graphics, Programming, Statistics, and Simulation
- reporttools: Generate LaTeX tables of descriptive statistics
- xtable: Export tables to LaTeX or HTML
- . . .
- For a general overview see CRAN Task View: Reproducible Research

R related materials

- Friedrich Leisch's The Sweave Homepage

- An Sweave Demo, Literate Programming in R, Reproducible Research

- Jeromy Anglim's talk about "R Workflow: Slides from a Talk at Melbourne R Users (1st Dec 2010)" (Slides + Video)

- Tal Galili: Exporting R output to MS-Word with R2wd (an example session)

- CRAN Task View: Reproducible Research

- A Sweave Wiki

- Customizing Sweave to Produce Better Looking LaTeX Output

Reproducible research related materials

- Hothorn/Leisch (2011): Case studies in reproducibility
- ReproducibleResearch.net
- stackoverflow: Workflow for statistical analysis and report writing
- Gentleman/Lang (2004): Statistical Analyses and Reproducible Research