Implementing the Compendium Concept with Sweave and DOCSTRIP

by Michael Lundholm

Abstract This article suggests an implementation of the compendium concept by combining Sweave and the LaTeX literate programming environment DOCSTRIP.

Introduction

Gentleman and Lang (2007) introduced *compendiums* as a mechanism to combine text, data, and auxiliary software into a distributable and executable unit, in order to achieve *reproducible research*[1]:

> "...research papers with accompanying software tools that allow the reader to directly reproduce the result and employ the methods that are presented ..." Gentleman and Lang (2007, (abstract))

Gentleman (2005) provides an example of how the compendium concept can be implemented. The core of the implementation is a Sweave[2] source file. This source file is then packaged together with data and auxiliary software as an R package.

In this article I suggest an alternative implementation of the compendium concept combining Sweave with DOCSTRIP. The latter is the LaTeX literate programming environment used in package documentation and as an installation tool.[3] DOCSTRIP has previously been mentioned in the context of reproducible research, but then mainly as a hard to use alternative to Sweave.[4] Here, instead, DOCSTRIP and Sweave are combined.

Apart from the possibility to enjoy the functionality of Sweave and packages such as **xtable** etc the main additional advantages are that

- in many applications almost all code and data can be kept in a single source file,
- multiple documents (i.e., PDF files) can share the same Sweave code chunks.

This means not only that administration of an empirical project is facilitated but also that it becomes easier to achieve reproducible research. Since DOCSTRIP is a part of every LaTeX installation a Sweave user need not install any additional software. Finally, Sweave and DOCSTRIP can be combined to produce more complex projects such as R packages.

One example of the suggested implemention will be given. It contains R code common to more than one document; an article containing the advertisement of the research (using the terminology of Buckheit and Donoho (1995)), and one technical documentation of the same research. In the following I assume that the details of Sweave are known to the readers of the R Journal. The rest of the article will (i) give a brief introduction to DOCSTRIP, (ii) present and comment the example and (iii) close with some final remarks.

DOCSTRIP

Suppose we have a source file the entire or partial content of which should be tangled into one or more result files. In order to determine which part of the source file that should be tangled into a certain result file (i) the content of the source file is tagged with none, one or more tags (tag–lists) and (ii) the various tag–lists are associated with the result files in a DOCSTRIP "installation" file.

There are several ways to tag parts of the source file:

- A single line: Start the line with '%<*tag-list*>'.
- Several lines, for instance one or more code or text chunks in Sweave terminology: On a single line before the first line of the chunk enter the start tag '%<*tag-list*>' and on a single line after the last line of the chunk the end tag '%</*tag-list*>'.
- All lines: Lines that should be in all result files are left untagged.

'tag-list' is a list of tags combined with the Boolean operators '|' (logical or), '&' (logical and) and '!' (logical negation). A frequent type of list would be, say, 'tag1|tag2|tag3' which will tangle the tagged material whenever 'tag1', 'tag2' *or* 'tag3' is called for into the result files these tags are associated with. The initial '%' of the tags must be in the file's first column or else the tag will not be recognised as a DOCSTRIP tag. Also, tags must be matched so a start tag with 'tag-list' must be closed by an end tag with 'tag-list'. This resembles the syntax of LaTeX environments rather than the Sweave syntax, where the end of a code or text chunk is indicated by the beginning of the next text or code chunk. Note also that tags cannot be nested.[5]

[1] Buckheit and Donoho (1995).
[2] Leisch. See also Leisch (2008) and Meredith and Racine (2008).
[3] Mittelbach et al. (2005) and Goossens et al. (1994, section 14.2).
[4] Hothorn (2006) and Rising (2008).
[5] More exactly: *Outer* tags, which are described here, cannot be nested but *inner tags* can be nested with outer tags. See Goossens et al. (1994, p. 820) for details.

The following source file ('docex.txt') exemplifies all three types of tags:

```
1  %<file1|file2>This line begins both files.
2  %<*file1>
3
4  This is the text that should be included in file1
5
6  %</file1>
7
8  This is the text to be included in both files
9
10 %<*file2>
11 This is the text that should be included in file2
12 %</file2>
13 %<*file1|file2>
14 Also text for both files.
15 %</file1|file2>
```

For instance, line 1 is a single line tagged '`file1`' *or* '`file2`', line 2 starts and line 6 ends a tag '`file1`' and line 13 starts and line 15 ends a tag '`file1`' *or* '`file2`'. Lines $7-9$ are untagged.

The next step is to construct a `DOCSTRIP` installation file which associates each tag with one or more result files:

```
1  \input docstrip.tex
2  \keepsilent
3  \askforoverwritefalse
4  \nopreamble
5  \nopostamble
6  \generate{
7  \file{file1.txt}{\from{docex.txt}{file1}}
8  \file{file2.txt}{\from{docex.txt}{file2}}
9  }
10 \endbatchfile
```

Line 1 loads `DOCSTRIP`. Lines $2-5$ contain options that basically tell `DOCSTRIP` not to issue any messages, to write over any existing result files and not to mess up the result files with pre– and post-ambles.[6] The action takes place on lines $6-9$ within the command '`\generate{}`', where lines $7-8$ associate the tags '`file1`' and '`file2`' in the source file 'docex.txt' with the result files 'file1.txt' and 'file2.txt'.[7]

We name this file 'docex.ins', where '`.ins`' is the conventional extension for `DOCSTRIP` installation files. `DOCSTRIP` is then invoked with

```
latex docex.ins
```

A log–file called 'docex.log' is created from which we here show the most important parts (lines $56-67$):

```
56 Generating file(s) ./file1.txt ./file2.txt
57 \openout0 = `./file1.txt'.
58
59 \openout1 = `./file2.txt'.
60
61
62 Processing file docex.txt (file1) -> file1.txt
63                           (file2) -> file2.txt
64 Lines  processed: 15
65 Comments removed: 0
66 Comments  passed: 0
67 Codelines passed: 8
```

We see that two result files are created from the 15 lines of code in the source file. First 'file1.txt';

```
1  This line begins both files.
2
3  This is the text that should be included in file1
4
5
6  This is the text to be included in both files
7
8  Also text for both files.
```

and 'file2.txt';

```
1  This line begins both files.
2
3  This is the text to be included in both files
4
5  This is the text that should be included in file2
6  Also text for both files.
```

Note that some lines are blank in both the original source file and the result files. Disregarding these the two result files together have 8 lines of code. The untagged material in lines $7-9$ in the source files is tangled into both result files, the blank lines 7 and 8 in the source file result in the blank lines 5 and 7 in 'file1.txt' and the blank lines 2 and 4 in 'file2.txt'.

Example

In the following a simple example will be given of how `DOCSTRIP` can be combined with `Sweave` to implement the compendium concept. The starting point is a "research problem" which involves loading some data into R, preprocessing the data, conducting an estimation and presenting the result. The purpose is to construct a single compendium source file which contains the code used to create (i) an "article" PDF–file which will provide a brief account of the test and (ii) a "technical documentation" PDF–file which gives a more detailed description of loading and preprocessing data and the estimation. The source file also contains the code of a BibTEX databse file and the

[6] Pre– and postambles are text lines that are starting with a comment character. Since result files may be processed by software using different comment characters some care is needed to use pre– and postambles constructed by `DOCSTRIP`. See Goossens et al. (1994, p. 829f and 833f) how to set up pre– and postambles that are common to all result files from a given installation file.

[7] From the example one infer that multiple source files are possible, altough the compendium implementation discussed later in most cases would have only one.

DOCSTRIP installation file. Although this source file is neither a LaTeX file or a Sweave file I will use the extension '.rnw' since it first run through Sweave. Here we simplify the example by using data from an R package, but if the data set is not too large it could be a part of the source file.

We can think of the "article" as the "advertisement" intended for journal publication and the "technical documentation" as a more complete account of the actual research intended to be available on (say) a web place. However, tables, graphs and individual statistics should originate from the same R code so whenever Sweave is executed these are updated in both documents. There may also be common text chunks and when they are changed in the source file, both documents are updated via the result files.

The example code in the file 'example_source.rnw' is as follows:

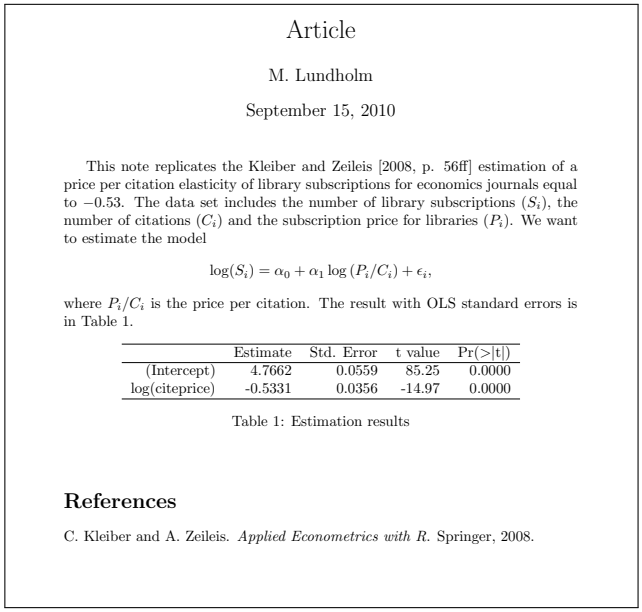
```
%<*article|techdoc>
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Author: Michael Lundholm
%% Email:  michael.lundholm@ne.su.se
%% Date:   2010-09-06
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% The project files consists of
%%  * example_source.rnw (THIS FILE)
%% To create other files execute
%%   R CMD Sweave example_source.rnw
%%   latex example.ins
%%   pdflatex example_techdoc.tex
%%   bibtex example_techdoc
%%   pdflatex example_article.tex
%%   bibtex example_article
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\documentclass{article}
\usepackage{Sweave,amsmath,natbib}
%</article|techdoc>
%<article>\title{Article}
%<techdoc>\title{Technical documentation}
%<*article|techdoc>
\author{M. Lundholm}
\begin{document}
\maketitle
This note replicates the \citet[p. 56ff]{Kleiber08}
estimation of a price per citation elasticity of
library subscriptions for economics journals equal
to $\input{coef.txt}$.
%</article|techdoc>
%<*techdoc>
The data, available in R package AER on CRAN, is loaded:
<<Loading data>>=
data("Journals",package="AER")
@
%</techdoc>
%<*article|techdoc>
The data set includes the number of library
subscriptions ($S_i$), the number of citations
($C_i$)  and the subscription price for libraries
($P_i$). We want to estimate the model
$$\log(S_i)=\alpha_0+\alpha_1
\log\left(P_i/C_i\right)+\epsilon_i,$$
where $P_i/C_i$ is the price per citation.
%</article|techdoc>
%<*techdoc>
We the define the price per citation, include
the variable in the data frame \texttt{Journals}
<<Define variable>>=
Journals$citeprice <- Journals$price/Journals$citations
@
and estimate the model:
<<Estimate>>=
result <- lm(log(subs)~log(citeprice),data=Journals)
@
%</techdoc>
%<*article|techdoc>
The result with OLS standard errors is in
Table~\ref{ta:est}.
<<Result,results=tex,echo=FALSE>>=
library(xtable)
xtable(summary(result),label="ta:est",
caption="Estimation results")
@
<<echo=FALSE>>=
write(round(coef(result)[[2]],2),file="coef.txt")
@
\bibliographystyle{abbrvnat}
\bibliography{example}
\end{document}
%</article|techdoc>
%<*bib>
 @Book{ Kleiber08,
author = {Christian Kleiber and Achim Zeileis},
publisher = {Springer},
year = {2008},
title = {Applied Econometrics with {R}}}
%</bib>
%<*dump>
<<Write DOCSTRIP installation file>>=
writeLines(
"\\input docstrip.tex
\\keepsilent
\\askforoverwritefalse
\\nopreamble
\\nopostamble
\\generate{
\\file{example_article.tex}%
{\\from{example_source.tex}{article}}
\\file{example_techdoc.tex}%
{\\from{example_source.tex}{techdoc}}
\\file{example.bib}%
{\\from{example_source.tex}{bib}}}
\\endbatchfile
",con="example.ins")
@
%</dump>
```

The compendium source file contains the following DOCSTRIP tags (for their association to files, see below):

- 'article' associated with 'example_article.tex', which contains the code to the "advertisement" article,
- 'techdoc' associated with 'example_techdoc.tex', which contains the code to the technical documentation,
- 'bib' associated with 'example.bib' which contains the code to the BibTeX data base file,
- 'dump' associated with no file.

Note that the tags 'article' and 'techdoc' overlap with eachother but not with 'bib' and 'dump', which in turn are mutually exclusive. There is no untagged material.

Lines 2 – 15 contain general information about the distributed project, which could be more or less elaborate. Here it just states that the project is distributed as a single source file and how the compendium source file should be processed to get the relevant output



Article

M. Lundholm

September 15, 2010

This note replicates the Kleiber and Zeileis [2008, p. 56ff] estimation of a price per citation elasticity of library subscriptions for economics journals equal to -0.53. The data set includes the number of library subscriptions (S_i), the number of citations (C_i) and the subscription price for libraries (P_i). We want to estimate the model

$$\log(S_i) = \alpha_0 + \alpha_1 \log(P_i/C_i) + \epsilon_i,$$

where P_i/C_i is the price per citation. The result with OLS standard errors is in Table 1.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.7662	0.0559	85.25	0.0000
log(citeprice)	-0.5331	0.0356	-14.97	0.0000

Table 1: Estimation results

References

C. Kleiber and A. Zeileis. *Applied Econometrics with R*. Springer, 2008.

Figure 1: 'example_article.pdf'.

'example_article.pdf' and 'example_techdoc.pdf'.

When the instructions are followed, `Sweave` is run first on 'example_source.rnw' creating the file 'example_source.tex', in which the `Sweave` code chunks are replaced by the corresponding R output code wrapped with LaTeX typesetting commands. One of the R functions used in this Sweave session is `writeLines()` (see the lines $80-96$) so that the `DOCSTRIP` installation file 'example.ins' is created before `DOCSTRIP` is run.

This file 'example_source.tex' is the `DOCSTRIP` source file from which the `DOCSTRIP` utility, together with the installation file 'example.ins', creates the result files 'example_article.tex', 'example_techdoc.tex' and 'example.bib'. The two first result files share some but not all code from the `DOCSTRIP` source file. The result files are then run with the LaTeX family of software (here `pdflatex` and BibTeX) to create two PDF–files 'example_article.pdf' and 'example_techdoc.pdf'. These are shown in Figures 1–2.

Note that the entire bibliography (BibTeX) file is included on lines $73-77$ and extracted with `DOCSTRIP`. Note also on line 73 that unless the @ indicating a new bibliographical entry is *not* in column 1 it is mixed up by `Sweave` as a new text chunk and will be removed, with errors as the result when BibTeX is run.[8]

The bibliography database file is common to both 'example_article.tex' and 'example_techdoc.tex'. Here the documents have the same single reference. But in real implementations bibliographies would probably not overlap completely. This way handling references is then preferable since all bibliographical references occur only once in the source file.[9]

In LaTeX cross references are handled by writing information to the auxiliary file, which is read by later LaTeX runs. This handles references to an object located both before and after the reference in the LaTeX file. In `Sweave` '' can be used to refer to R objects created before but not after the reference is made. This is not exemplified here. But since `Sweave` and LaTeX are run sequentially an object can be created by R, written to a file (see the code chunk on lines $65-67$) and then be used in the LaTeX run with the command '`\input{}`' (see code line 29).

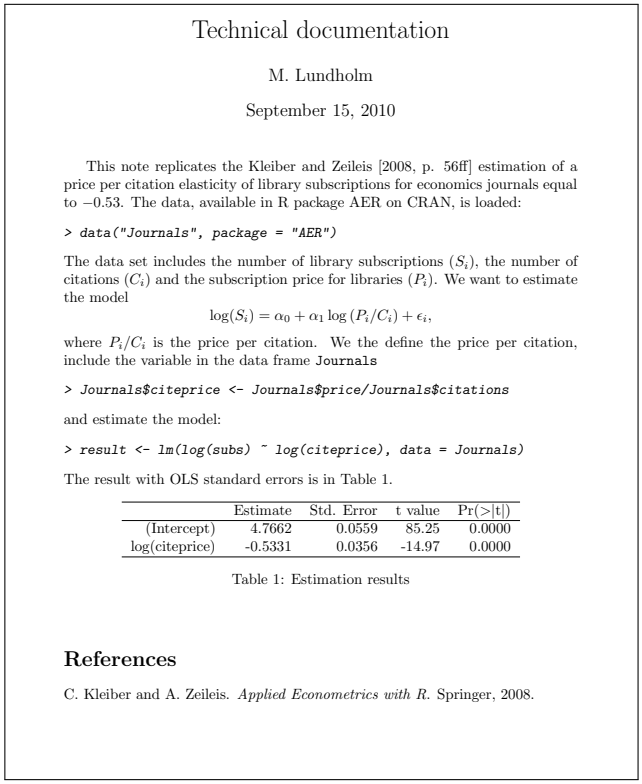
Final comments

By making use of combinations of `DOCSTRIP` and (say) '`writeLines()`' and by changing the order in which `Sweave` and `DOCSTRIP` are executed the applications can be made more complex. Such examples may be found Lundholm (2010a,b).[10] Also, the use of `DOCSTRIP` can facilitate the creation of R packages as exemplified by the R data package **sifds** available on CRAN (Lundholm, 2010c). Another type of example would be teaching material, where this article may itself serve as an example. Apart from the `DOCSTRIP`

[8] The tag '`dump`' is a safeguard against that this material is allocated to some result file by `DOCSTRIP`; in this case to the BibTeX data base file.

[9] One alternative would be to replace the command '`\bibliography{example}`' on line 69 with the content of 'example_article.bbl' and 'example_techdoc.bbl' appropriately tagged for `DOCSTRIP`. However, this procedure would require an "external" bibliography data base file. The problem then is that each time the data base is changed, manual updating of the parts of 'example_source.rnw' that creates 'example_article.bbl' and 'example_techdoc.bbl' is required. Creating the bibliography data base file via `DOCSTRIP` makes this manual updating unnecessary.

[10] An early attempt to implement the ideas presented in this article can be found in Arai et al. (2009).



Technical documentation

M. Lundholm

September 15, 2010

This note replicates the Kleiber and Zeileis [2008, p. 56ff] estimation of a price per citation elasticity of library subscriptions for economics journals equal to −0.53. The data, available in R package AER on CRAN, is loaded:

```
> data("Journals", package = "AER")
```

The data set includes the number of library subscriptions (S_i), the number of citations (C_i) and the subscription price for libraries (P_i). We want to estimate the model

$$\log(S_i) = \alpha_0 + \alpha_1 \log(P_i/C_i) + \epsilon_i,$$

where P_i/C_i is the price per citation. We the define the price per citation, include the variable in the data frame `Journals`

```
> Journals$citeprice <- Journals$price/Journals$citations
```

and estimate the model:

```
> result <- lm(log(subs) ~ log(citeprice), data = Journals)
```

The result with OLS standard errors is in Table 1.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.7662	0.0559	85.25	0.0000
log(citeprice)	-0.5331	0.0356	-14.97	0.0000

Table 1: Estimation results

References

C. Kleiber and A. Zeileis. *Applied Econometrics with R*. Springer, 2008.

Figure 2: 'example_techdoc.pdf'.

installation file and a Bash script file all code used to produce this article is contained in a single source file. The Bash script, together with DOCSTRIP, creates all example files including the final PDF–files; that is, all example code is executed every time this article is updated. So, if the examples are changed an update of the article via the Bash script also updates the final PDF–files in Figures 1–2.[11]

Colophone

This article was written on a i486-pc-linux-gnu platform using R version 2.11.1 (2010-05-31), LaTeX2$_\varepsilon$ (2005/12/01) and DOCSTRIP 2.5d (2005/07/29).

Acknowledgement

The compendium implementation presented here is partially developed in projects joint with Mahmood Arai, to whome I am owe several constructive comments on a previous version.

Bibliography

M. Arai, J. Karlsson, and M. Lundholm. On fragile grounds: A replication of *Are Muslim immigrants different in terms of cultural integration?* Accepted for publication in the Journal of the European Economic Association, 2009. URL http://www.eeassoc.org/index.php?site=JEEA&page=55.

J. B. Buckheit and D. L. Donoho. WaveLab and reproducible research. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and statistics*, Lecture notes in statistics 103, pages 55–81. Springer Verlag, 1995.

R. Gentleman. Reproducible research: A bioinformatics case study. 2005. URL http://www.bioconductor.org/docs/papers/2003/Compendium/Golub.pdf.

R. Gentleman and D. T. Lang. Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, 16:1–23, 2007. URL http://pubs.amstat.org/doi/pdfplus/10.1198/106186007X178663.

M. Goossens, F. Mittelbach, and A. Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, MA, USA, second edition, 1994.

T. Hothorn. Praktische aspekte der reproduzierbarkeit statistischer analysen in klinischen studien. Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2006. URL http://www.imbe.med.uni-erlangen.de/~hothorn/talks/AV.pdf.

[11] The compendium source files of projects mentioned in this paragraph, including this article, can be found at http://people.su.se/~lundh/projects/.

F. Leisch. Sweave: Dynamic generation of statistical reports using literate data analysis. In W. Härdle and y. . . Bernd Rönz, pages = 575–580, editors, *Compstat 2002 – Proceedings in Computational Statistics*.

F. Leisch. *Sweave User Manual*, 2008. URL http://www.stat.uni-muenchen.de/~leisch/Sweave/Sweave-manual.pdf. R version 2.7.1.

M. Lundholm. Are inflation forecasts from major swedish forecasters biased? Research paper in economics 2010:10, Department of Economics, Stockholm University, 2010a. URL http://swopec.hhs.se/sunrpe/abs/sunrpe2010_0010.htm.

M. Lundholm. Sveriges riksbank's inflation interval forecasts 1999–2005. Research paper in economics 2010:11, Department of Economics, Stockholm University, 2010b. URL http://swopec.hhs.se/sunrpe/abs/sunrpe2010_0010.htm.

M. Lundholm. *sifds: Swedish inflation forecast data set*, 2010c. URL http://www.cran.r-project.org/web/packages/sifds/index.html. R package version 0.9.

E. Meredith and J. S. Racine. Towards reproducible econometric research: The Sweave framework. *Journal of Applied Econometrics*, 2008. URL http://dx.doi.org/10.1002/jae.1030. Published Online: 12 Nov 2008.

F. Mittelbach, D. Duchier, J. Braams, M. Wolińsky, and M. Wooding. The DOCSTRIP program. Version 2.5d, 2005. URL http://tug.ctan.org/tex-archive/macros/latex/base/.

B. Rising. Reproducible research: Weaving with Stata. StataCorp LP, 2008. URL http://www.stata.com/meeting/italy08/rising_2008.pdf.

Michael Lundholm
Department of Economics
Stockholm University
SE–106 91 Stockholm
Sweden
michael.lundholm@ne.su.se