



TOWARDS REPRODUCIBLE ECONOMETRIC RESEARCH: THE SWEAVE FRAMEWORK

EVAN MEREDITH AND JEFFREY S. RACINE*

Department of Economics, McMaster University, Hamilton, Ontario, Canada

1. OVERVIEW

The Sweave package for the R and S-plus statistical computing environments enables the user to construct a single file which includes both the code to be run in R/S-plus and the $\text{T}_{\text{E}}\text{X}/\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$ code comprising the text of the document. Files containing both types of code are referred to as `.Rnw/.Snw` files. The various sections ('chunks') of R/S-plus and $\text{T}_{\text{E}}\text{X}/\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$ code are included in the file in the order in which they are to be employed in the final document. Sweave then weaves together the code chunks to produce a `.tex` file that may be compiled using TeX/LaTeX. By using Sweave, an individual can create a dynamic document (Gentleman and Lang, 2004) which includes both the statistical analysis and the methods by which the output underlying the analysis is obtained. This process sidesteps a major source of research errors, namely, the misreporting of computer output.

Sweave was developed by Friedrich Leisch and is written in the S language.¹ A directory containing the manual (Leisch, 2006), a number of other related documents (Leisch, 2002a,b, 2003a,b) and a FAQ can be found at <http://www.ci.tuwien.ac.at/~leisch/Sweave/>. Utilizing Sweave requires R/S-plus² and an implementation of $\text{T}_{\text{E}}\text{X}/\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$. In what follows, we restrict attention to the open R (Team, 2008) implementation of the S language and $\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$. There is no need to install Sweave itself, as it is included in the `utils` package in R version 1.5.0 or higher. Windows users should avoid installing R in the default installation directory, Program Files, as the installation path contains blank spaces which $\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$ cannot handle. To remedy this, simply install R in a path that contains no blanks (Leisch, 2006).

* Correspondence to: Jeffrey S. Racine, Department of Economics, Kenneth Taylor Hall, McMaster University, Hamilton, Ontario, Canada L8S 4M4. E-mail: racinej@mcmaster.ca

¹ S (Chambers and Becker, 1984) is a statistical computing language designed at Bell Laboratories.

² Download information and documentation for the open R (Team, 2008) and commercial S-plus implementations of the S language are located at <http://www.r-project.org/> and <http://www.insightful.com/>, respectively.

The main strength of *Sweave* is the ease with which it makes reproducible research possible. Unlike the reproduction of experimental research, which may require the replication of experimental results, to reproduce the results of econometric and statistical analysis, one only needs access to the data and the code employed in the computations. With *Sweave*, anyone in possession of the *.Rnw* file, the data, and the relevant software may reproduce the results using the methods the original author employed. Any individuals who question the validity of the output may satisfy their curiosity by using *Sweave* on the *.Rnw* file themselves. Furthermore, the document is itself interactive, in that the R code chunks contained within may be manipulated. For example, estimation techniques or functional forms may be changed and then the computations rerun using the same data and other relevant code included in the original analysis. Supplying *.Rnw* files to students based on materials presented during lectures may also allow for an enhanced learning experience due to this flexibility (Vinod, 2001).

Sweave also contains a very flexible environment for the original authors to modify their estimation and analysis. When using a word processor in conjunction with some statistical software, changing the empirical strategy involves changing code, running the code, and then cutting and pasting to the word processor while making the necessary adjustments to the final document. With *Sweave*, only the *.Rnw* file needs to be changed, and, since *Sweave* embeds the output of R, only small changes need to be made to the text to modify the final document. This feature allows the same statistical methods to be used with a different dataset or different methods to be used with the same dataset with minimal changes made to the existing code in the *.Rnw* file. This also prevents any confusion over which code was utilized to produce the results contained in the final document, as they are both contained in the same file.

The remainder of this article discusses *Sweave*'s relation to the concept of 'Literate Programming' as developed by Don Knuth (1992), details how one constructs a *.Rnw* file using code chunks, and then works through some examples to illustrate the features and flexibility of *Sweave*.

2. LITERATE PROGRAMMING AND REPRODUCIBLE RESEARCH

The main points underlying the concept of Literate Programming may be summarized as follows (Greyer, 2006):

1. Programs are useless without descriptions.
2. Descriptions should be literate, not comments in code or typical reference manuals.
3. The code in the descriptions should work. Thus it is necessary to extract the real working code from the literary description.

The following quotation from Don Knuth (1992) is also illuminating:

Let us change our traditional attitude to the construction of programs. Instead of imagining that our main task is to instruct a *computer* what to do, let us concentrate rather on explaining to *human beings* what we want a computer to do.

Sweave embraces these ideas. The *.Rnw* file contains the text of the document, which describes the methods employed and the analysis undertaken, and the code, which tells a computer how to

implement the methods outlined in the text. The program therefore contains literate descriptions of the code used that are, hopefully, less nebulous and more literate than typical reference manual descriptions or code comments. By weaving the `.Rnw` file, it can easily be seen whether the code contained in the descriptions works, while the working code can be extracted from the `.Rnw` file through the use of the `Stangle` command.

Reproducibility is also aided through the use of open source or non-proprietary software. It has been noted that the use of proprietary software, which requires the user to purchase the program, presents an obstacle to reproducibility that open source software, which everyone with the relevant OS has access to free of charge, does not (de Leeuw, 2001; Baiocchi, 2007). The programs involved in the use of `Sweave`, if using R, are open source and are available to users across a number of platforms including Microsoft Windows, Apple's MacOS, several GNU/Linux distributions, and various variants of UNIX such as BSD.

3. WRITING `.Rnw` FILES

In order to create a document to `Sweave`, the user must write a `.Rnw` file. A `.Rnw` file contains the various \LaTeX and R code chunks arranged in the order in which they are to be employed in the weaving of the final document. To write such a file, the NOWEB (Ramsey, 2008) syntax is used.

With the NOWEB syntax, code chunks begin with one of the following commands. `<<>>=` denotes the start of an R code chunk, and `@` denotes the start of a \LaTeX code chunk. By default, the first code chunk is a \LaTeX one, so there is no need to insert `@` at the start of your `.Rnw` file if the document is to start with text first. If the user wishes to use \LaTeX syntax, the option is also available.

The `<<>>=` command allows for several options, separated by commas, to be inserted between `<<` and `>>`. By default the first of these is the name of the code chunk. Following this, a number of other options may be inserted, resulting in the general form `<< name, options >> =`. The available options and their default settings are presented in Table I.

3.1. Weaving and Tangling

Once the `.Rnw` file has been written, two basic operations can be performed upon it: weaving and tangling.³ Weaving evaluates the R code chunks and then produces a `.tex` file containing the text and R code chunks in the order in which they appeared in the `.Rnw` file and executed with the various options included in the precursor to the R code chunks. Tangling, on the other hand, extracts the R chunks from the `.Rnw` file, producing a file containing the code only.

To weave a `.Rnw` file within the R terminal or GUI, use the following command: `Sweave('filename.Rnw')`. The relevant code chunks are then evaluated in R, and a `.tex` file is produced which can be used with \LaTeX to produce the final document as desired.

To tangle a `.Rnw` file, in order to extract the R code contained within it, simply use the following command within the R terminal or GUI: `Stangle('filename.Rnw')`. The `Stangle` command may also include a number of optional arguments; for details, see the `Sweave` manual

³ Weaving and tangling are operations developed by Don Knuth in conjunction with his concept of Literate Programming (Knuth, 1992).

Table I. R Code chunk options

Option	Description	Default
echo	If TRUE, the R code is included in the output file. If FALSE, it is not	TRUE
eval	If TRUE, the code included in the chunk is evaluated. If FALSE, it is not	TRUE
results	Indicates the format of the output. If <code>verbatim</code> , it is included in an R like output. If <code>tex</code> , it is taken as properly coded \LaTeX output. If <code>hide</code> , it is not included in the output (though the code is executed)	<code>verbatim</code>
term	If TRUE, values of assignments are not printed, values of single objects are printed. If FALSE, output is only from <code>print</code> or <code>cat</code> statements	FALSE
print	If TRUE, each expression in the code chunk is wrapped into a <code>print()</code> statement before evaluation; the values of all expressions become visible	FALSE
split	If TRUE, text output is written to separate files for each code chunk	FALSE
strip.white	If TRUE, blank lines at the start and end of each output are removed If <code>all</code> , then all blank lines are removed	FALSE
prefix	If TRUE, generated filenames of figures and output have a common prefix	TRUE
prefix.string	A character string, the default of which is the name of the <code>.Rnw</code> file	
include	Indicates whether input statements for text output and <code>include-graphics</code> statements for figures should be auto-generated	TRUE
fig	Indicates if the code chunk to be evaluated produces graphical output. Only one figure per code chunk is allowed	FALSE
eps	Indicates whether EPS figures are generated	TRUE
pdf	Indicates whether PDF figures are generated	TRUE
width	Numeric argument for width of figures in inches	6
height	Numeric argument for height of figures in inches	6

(Leisch, 2006). This produces a file with the extension `.R` containing the R code chunks only. By default, the extracted chunks are separated by comments that include the names and numbers of the various code chunks included in the `.Rnw` file.

4. CONSTRUCTING A `.Rnw` FILE: AN ILLUSTRATIVE EXAMPLE

To demonstrate how to use *Sweave* to create a dynamic document and also to highlight its flexibility, the *Bwages* dataset from the *Ecdat* package available in R is used.⁴ The dataset contains 1472 individual cross-sectional observations on gross hourly wage rate in euros (*wage*), education level from 1 [low] to 5 [high] (*educ*), years of experience (*exper*), and a gender indicator (*sex*). The following bit of code, as it would appear in a `.Rnw` file, formats the dataset to be used in the following examples:

```
<<loaddata, echo=FALSE, results=hide>>=
#load the Ecdat package
library(Ecdat)
#load the Bwages dataset
```

⁴ In order to run this example you must have the *Ecdat* package installed as it is not part of the base R distribution.

```
data(Bwages)
#build the dataset to be used in the example, data.ex
#drop the sex variable from the data set
dataex <- Bwages[,-4]
```

The options selected in `<<>>=` are set so that the R code and output do not appear in the final .tex file.

The following bit of code runs a simple least-squares linear regression of wages on educ and exper:

```
<<modell>>=
#define and estimate the linear model, modell, using dataex
modell <- lm(wage~educ + exper, data=dataex)
#report the results of the estimation of modell
summary(modell)
```

Using Sweave on such a code chunk would produce the following output in the final document:

```
> modell <- lm(wage~educ + exper, data = dataex)
> summary(modell)
```

Call:

```
lm(formula = wage~educ + exper, data = dataex)
```

Residuals:

```
Min 1Q Median 3Q Max
-14.0436 -2.0808 -0.4068 1.5915 31.2307
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.07374 0.37269 2.881 0.00402 **
educ 1.93037 0.08154 23.674 < 2e-16 ***
exper 0.20069 0.00966 20.774 < 2e-16 ***
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 3.606 on 1469 degrees of freedom

Multiple R-squared: 0.3445, Adjusted R-squared: 0.3436

F-statistic: 386 on 2 and 1469 DF, p-value: < 2.2e-16

The `<<>>=` that began the code chunk contained no other entries besides a name for the code chunk. Therefore, Sweave operated with all options at default settings; hence the code will be repeated in the final text, and all output will appear as it would appear in R. It may be the case

that, in the final document, the author does not wish for the code to be repeated. To specify this option, simply set `echo=FALSE` within `<<>>=`. It may also be the case that the author wishes the results of the regression to appear in the final document in a more visually appealing format. An excellent tool for such an operation is the `xtable` command contained in the `xtable` package (Dahl, 2004). `xtable` creates a table in \LaTeX format automatically for certain classes of R functions, including the `lm` function used above. To use `xtable` to create a table of the results of the regression above, use the following code:

```
<<modelltable, echo=FALSE, results=tex>>=
#load the xtable package
library(xtable)
#create a table for modell
xtable(modell, caption="modell Estimation Output")
```

The options specified in this code chunk suppress the code from the final document and format the output as \LaTeX code; this requires that the option `results=tex` be used. This generates the results shown in Table II.

Through the use of `Sweave` and `xtable`, the desired regressions can be coded, evaluated, and reported automatically every time `Sweave` is run on the `.Rnw` file. There is no need to return to the statistical software package you may use otherwise and then cut and paste the results back into a word processor while making modifications to the original tables.

Now suppose that, after running `Sweave` on your `.Rnw` file, you decide that some aspect of your original analysis was lacking. In the example considered here, suppose that you now wish to regress the log of wages on education and a quadratic experience term. Let this model be called `modell2` to keep it separate from the earlier model. This modification can be made using `Sweave` through a trivial change in the relevant code chunks in the `.Rnw` file. Compare the following code, which runs the new regression and generates a table of the results, to the original code chunks to see the small changes that are required to make this adjustment to the final document.

```
<<modell2, echo=FALSE, results=hide>>=
#define and estimate the model, modell2, using dataex
modell2 <- lm(log(wage)~educ + exper +I(exper^ 2), data=dataex)
#report the results of the estimation of modell2
summary(modell2)
@
<<modell2table, echo=FALSE, results=tex>>=
```

Table II. modell estimation output

	Estimate	SE	<i>t</i> -value	Pr(> <i>t</i>)
(Intercept)	1.0737	0.3727	2.88	0.0040
educ	1.9304	0.0815	23.67	0.0000
exper	0.2007	0.0097	20.77	0.0000

```
#create a table for model2
xtable(model2, caption="model2 Estimation Output")
```

Evaluating these code chunks would produce the results shown in Table III in the final document:

The changes can be made to the final document with minimal modification of the .Rnw file and no cutting and pasting of results between programs.

Figures can be included in the woven document by constructing a code chunk and using the option `fig=TRUE`. With the default settings, only one figure can be evaluated per code chunk. In order to include multiple figures, the user must specify some hook function to be executed before the code chunk is evaluated. For each logical option of a code chunk, a hook function can be specified and then executed if the option is set to `TRUE`. The following code specifies a figure hook to allow for the inclusion of four figures and includes the figures produced by the `plot(model2)` command in the final document:

```
<<multifig, results=hide>>=
options(SweaveHooks = list(multifig = function()
  par(mfrow=c(2,2))))
@

\begin{figure}
\begin{center}

<<model2, fig=TRUE, multifig=TRUE, echo=FALSE>>=
plot(model2)
@

\captionPlots produced with \texttt{plot(model2)}

\end{center}
\end{figure}
```

This code chunk yields Figure 1.

Another useful command is `\Sexpr`, which can be used to call R expressions within the body of text code chunks. For example, the command `\Sexprnrow(dataex)` can be used to call the number of observations in the dataset in the following sentence. The dataset `dataex` contains 1472 observations. `\Sexpr` can also be used in the construction of tables when the `xtable`

Table III. model2 estimation results

	Estimate	SE	<i>t</i> -value	Pr(> <i>t</i>)
(Intercept)	1.3867	0.0339	40.86	0.0000
educ	0.1596	0.0065	24.62	0.0000
exper	0.0352	0.0026	13.66	0.0000
I(exper ²)	-0.0005	0.0001	-7.63	0.0000

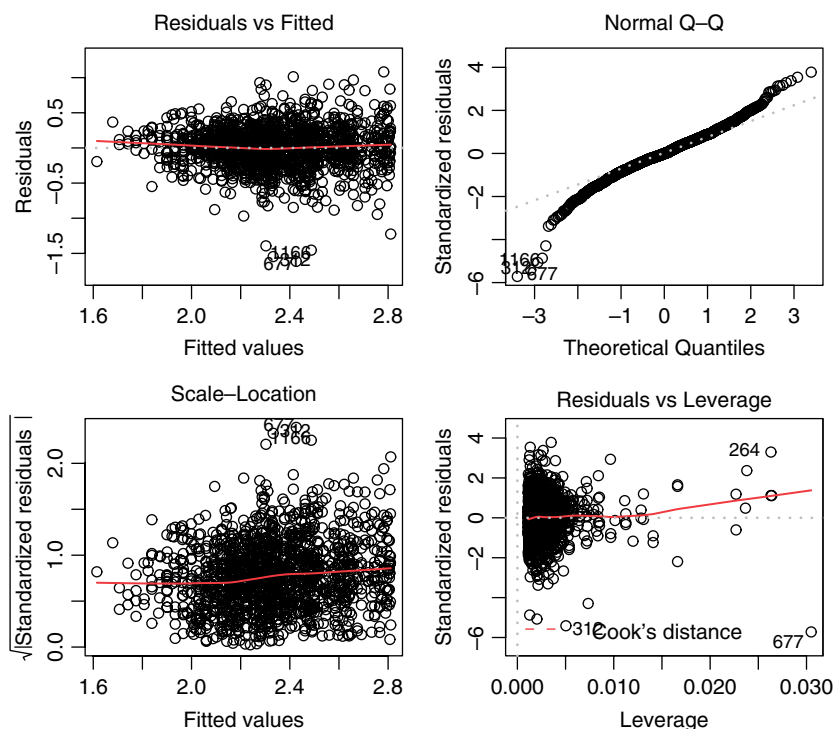


Figure 1. Plots produced by `plot(model2)`. This figure is available in color online at www.interscience.wiley.com/journal/jae

command may not be applicable. The table can be constructed manually, and then `\Sexpr` can be used to automatically fill in the elements of the table with the relevant R output. Any characters are allowed to be present in the expression contained within `\Sexpr` except for `{or}`. This is not a problem, as any complicated R expressions may be included in a hidden R code chunk earlier and then called using `\Sexpr` (Leisch, 2006); `\Sexpr` can call upon any computations computed earlier in the `.Rnw` file that were not set to be ignored.

For those researchers conducting computationally expensive and time-consuming operations, the `weaver` package (Falcon, 2006) may be of some use. `weaver` enables code chunks contained within the `.Rnw` file to be cached. This can be of use when certain code chunks require a large amount of time to be evaluated, and the author needs to make other changes to the document which don't affect the time-consuming chunks. Using `Sweave` on the document evaluates all the code chunks and takes as much time as running the entire code through R on its own. With `weaver`, the `.Rnw` file may be written so that the more time-consuming codes are cached when initially running `Sweave` so that other changes, not affecting them, may be made without reevaluating the entire code.

5. SUMMARY

`Sweave` provides a framework for mixing computer code such as R/S-plus and narrative text such as $\text{T}_{\text{E}}\text{X}/\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$ in a single document. This provides a research environment that sidesteps

a major source of research errors, namely, the misreporting of computer output. This framework streamlines much of the drudgery associated with research projects, such as the manual creation of tables, figures, and the like, and allows the researcher to instead concentrate on the research itself. Sweave is an ideal framework for students and faculty alike, and it helps migrate researchers towards reproducible econometric research.

ACKNOWLEDGEMENTS

The authors would like to thank James MacKinnon for his guidance and encouragement. All errors remain, of course, our responsibility.

REFERENCES

- Baiocchi G. 2007. Reproducible research in computational economics: guidelines, integrated approaches and open source software. *Computational Economics* **30**: 19–40.
- Chambers J, Becker R. 1984. *S: An Interactive Environment for Data Analysis and Graphics*. Wadsworth and Brooks/Cole: Pacific Grove, CA.
- Dahl DB. 2004. *The xtable Package*. <http://roadrunner.cancer.med.umich.edu/comp/docs/R/xtable.pdf> [16 September 2008].
- de Leeuw J. 2001. Reproducible research: the bottom line. Department of Statistics Papers 2001031101, Department of Statistics, UCLA.
- Falcon S. 2006. *How to Use Weaver for Sweave Document Processing*. http://www.bioconductor.org/packages/2.0/bioc/vignettes/weaver/inst/doc/weaver_howTo.pdf [16 September 2008].
- Gentleman R, Lang DT. 2004. Statistical analyses and reproducible research. Bioconductor Project Working Papers paper 2, Bioconductor Project.
- Greyer C. 2006. Sweave demo web site. <http://www.stat.umn.edu/~charlie/Sweave/> [16 September 2008].
- Knuth D. 1992. *Literate Programming*. Center for the Study of Language and Information: Stanford, CA.
- Leisch F. 2002a. Sweave: dynamic generation of statistical reports using literate data analysis. In *Compstat 2002: Proceedings in Computational Statistics*, Hardle W, Ronz, B (eds). 575–580.
- Leisch F. 2002b. Sweave, part I: Mixing R and \LaTeX . *R News* **2**(3): 28–31.
- Leisch F. 2003a. Sweave and beyond: computations on text documents. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)* Hornik K, Leisch F, Zeileis A (eds).
- Leisch F. 2003b. Sweave, part ii: package vignettes. *R News* **3**(2): 21–24.
- Leisch F. 2006. *Sweave Users Manual*. <http://www.statistik.lmu.de/~leisch/Sweave/Sweave-manual.pdf> [05 December 2008].
- Ramsey N. 2008. Noweb: a simple, extensible tool for literate programming. <http://www.eecs.harvard.edu/nr/noweb/> [16 September 2008].
- Team RDC. 2008. R: a language and environment for statistical computing. <http://www.r-project.org/> [16 September 2008].
- Vinod H. 2001. Care and feeding of reproducible econometrics. *Journal of Econometrics* **100**: 87–88.