The nuts and bolts of Sweave/Knitr for reproducible research

Marcus W. Beck

ORISE Post-doc Fellow
USEPA NHEERL Gulf Ecology Division, Gulf Breeze, FL
Email: beck.marcus@epa.gov, Phone: 850 934 2480

January 15, 2014

Reproducible research

In it's most general sense... the ability to reproduce results from an experiment or analysis conducted by another.

Reproducible research

In it's most general sense... the ability to reproduce results from an experiment or analysis conducted by another.

From Wikipedia... 'The ultimate product is the paper along with the full computational environment used to produce the results in the paper such as the code, data, etc. that can be used to reproduce the results and create new work based on the research.'

Reproducible research

In it's most general sense... the ability to reproduce results from an experiment or analysis conducted by another.

From Wikipedia... 'The ultimate product is the paper along with the full computational environment used to produce the results in the paper such as the code, data, etc. that can be used to reproduce the results and create new work based on the research.'

Concept is strongly based on the idea of literate programming such that the logic of the analysis is clearly represented in the final product by combining computer code/programs with ordinary human language [Knuth, 1992].

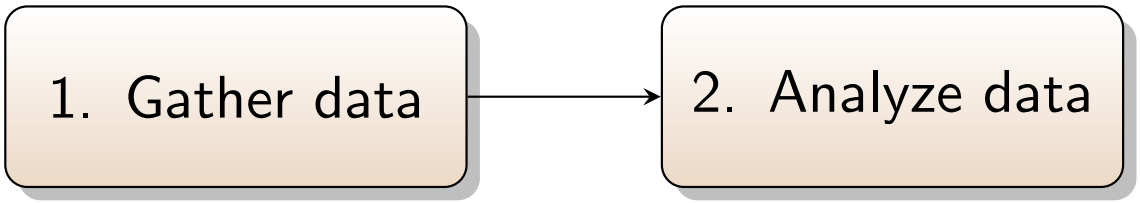
Non-reproducible research

Non-reproducible research

1. Gather data

- Begins with general question or research objectives
- Data collected in raw format (hard copy) converted to digital (Excel spreadsheet)

Non-reproducible research



- Begins with general question or research objectives
- Data collected in raw format (hard copy) converted to digital (Excel spreadsheet)

- Import data into stats program or analyze directly in Excel
- Create figures/tables directly in stats program
- Save relevant output

Non-reproducible research

- Begins with general question or research objectives
- Data collected in raw format (hard copy) converted to digital (Excel spreadsheet)

- Import data into stats program or analyze directly in Excel
- Create figures/tables directly in stats program
- Save relevant output

- Create research report using Word or other software
- Manually insert results into report
- Change final report by hand if methods/analysis altered

Reproducible research

- Begins with general question or research objectives
- Data collected in raw format (hard copy) converted to digital (text file)

- Create integrated script for importing data (data path is known)
- Create figures/tables directly in stats program
- No need to export (reproduced on the fly)

- Create research report using RR software
- Automatically include results into report
- Change final report automatically if methods/analysis altered

Reproducible research in R

Easily adopted using RStudio [http://www.rstudio.com/]

Also possible w/ Tinn-R or via command prompt but not as intuitive

Requires a LaTeX distribution system - use MikTex for Windows [http://miktex.org/]

Reproducible research in R

Easily adopted using RStudio [http://www.rstudio.com/]

Also possible w/ Tinn-R or via command prompt but not as intuitive

Requires a LaTeX distribution system - use MikTex for Windows [http://miktex.org/]

Essentially a LaTeX document that incorporates R code...

Uses Sweave (or Knitr) to convert .Rnw file to .tex file, then LaTeX to create pdf

Sweave comes with `utils` package, may have to tell R where it is

Reproducible research in R

Use same procedure for compiling a LaTeX document with one additional step

Reproducible research in R

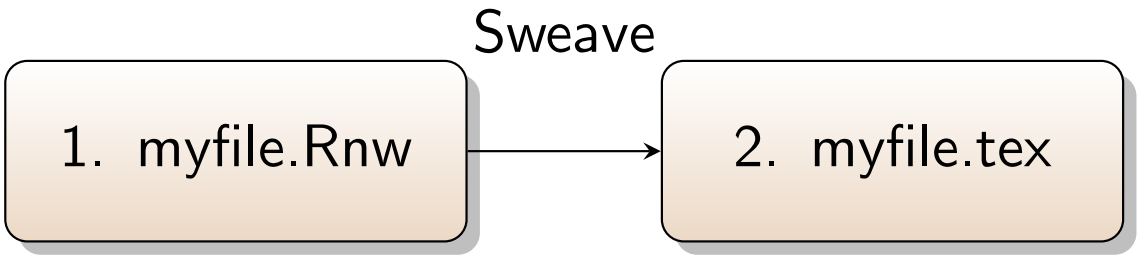
Use same procedure for compiling a LaTeX document with one additional step

1. myfile.Rnw

- A .tex file but with .Rnw extension
- Includes R code as 'chunks' or inline expressions

Reproducible research in R

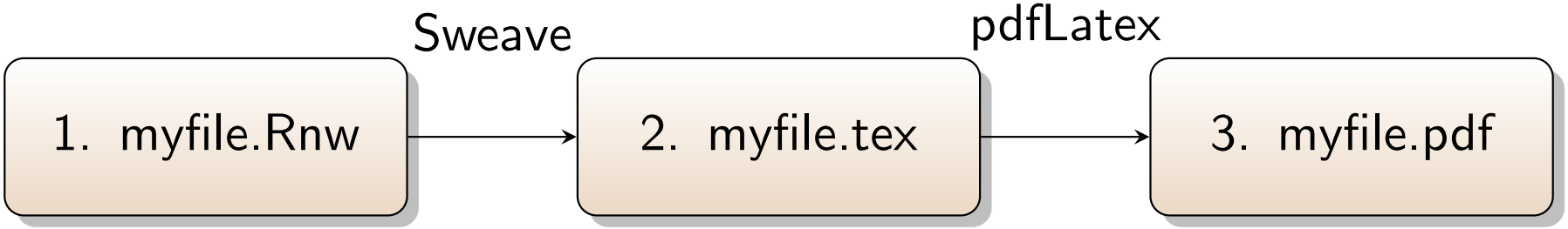
Use same procedure for compiling a LaTeX document with one additional step



- A .tex file but with .Rnw extension
- Includes R code as 'chunks' or inline expressions
- .Rnw file is converted to a .tex file using Sweave
- .tex file contains output from R, no raw R code

Reproducible research in R

Use same procedure for compiling a LaTeX document with one additional step



- A .tex file but with .Rnw extension
- Includes R code as 'chunks' or inline expressions

- .Rnw file is converted to a .tex file using Sweave
- .tex file contains output from R, no raw R code

- .tex file converted to pdf (or other output) for final format
- Include biblio with bibtex

Reproducible research in R

.Rnw file

```
\documentclass{article}
\usepackage{Sweave}

\begin{document}

Here's some R code:

<<eval=true,echo=true>>=
options(width=60)
set.seed(2)
rnorm(10)
@

\end{document}
```

Reproducible research in R

.tex file

```
\documentclass{article}
\usepackage{Sweave}

\begin{document}

Here's some R code:

\begin{Schunk}
\begin{Sinput}
> options(width=60)
> set.seed(2)
> rnorm(10)
\end{Sinput}
\begin{Soutput}
 [1] -0.89691455  0.18484918  1.58784533 -1.13037567
 [5] -0.08025176  0.13242028  0.70795473 -0.23969802
 [9]  1.98447394 -0.13878701
\end{Soutput}
\end{Schunk}

\end{document}
```

Reproducible research in R

The final product:

Here's some R code:

```
> options(width=60)
> set.seed(2)
> rnorm(10)

 [1] -0.89691455  0.18484918  1.58784533 -1.13037567
 [5] -0.08025176  0.13242028  0.70795473 -0.23969802
 [9]  1.98447394 -0.13878701
```

Sweave - code chunks

R code is entered in the LaTeX document using 'code chunks'

```
<<>>=
@
```

Any text within the code chunk is interpreted as R code

Arguments for the code chunk are entered within `<<here>>`

Sweave - code chunks

R code is entered in the LaTeX document using 'code chunks'

```
<<>>=
@
```

Any text within the code chunk is interpreted as R code

Arguments for the code chunk are entered within `<<here>>`

- `eval`: evaluate code, default `T`
- `echo`: return source code, default `T`
- `results`: format of output (chr string), default is 'include' (also 'tex' for tables or 'hide' to suppress)
- `fig`: for creating figures, default `F`

Sweave - code chunks

Changing the default arguments for the code chunk:

```
<<>>=
2+2
@
```

```
> 2+2
```

```
[1] 4
```

Sweave - code chunks

Changing the default arguments for the code chunk:

```
<<>>=
2+2
@
```

```
> 2+2

[1] 4
```

```
<<eval=F,echo=F>>=
2+2
@
```

Returns nothing...

Sweave - code chunks

Changing the default arguments for the code chunk:

```
<<>>=
2+2
@
```

```
> 2+2
[1] 4
```

```
<<eval=F>>=
2+2
@
```

```
> 2+2
```

```
<<eval=F,echo=F>>=
2+2
@
```

Returns nothing...

Sweave - code chunks

Changing the default arguments for the code chunk:

```
<<>>=
2+2
@
```

```
> 2+2
```

```
[1] 4
```

```
<<eval=F,echo=F>>=
2+2
@
```

Returns nothing...

```
<<eval=F>>=
2+2
@
```

```
> 2+2
```

```
<<echo=F>>=
2+2
@
```

```
[1] 4
```

Sweave - figures

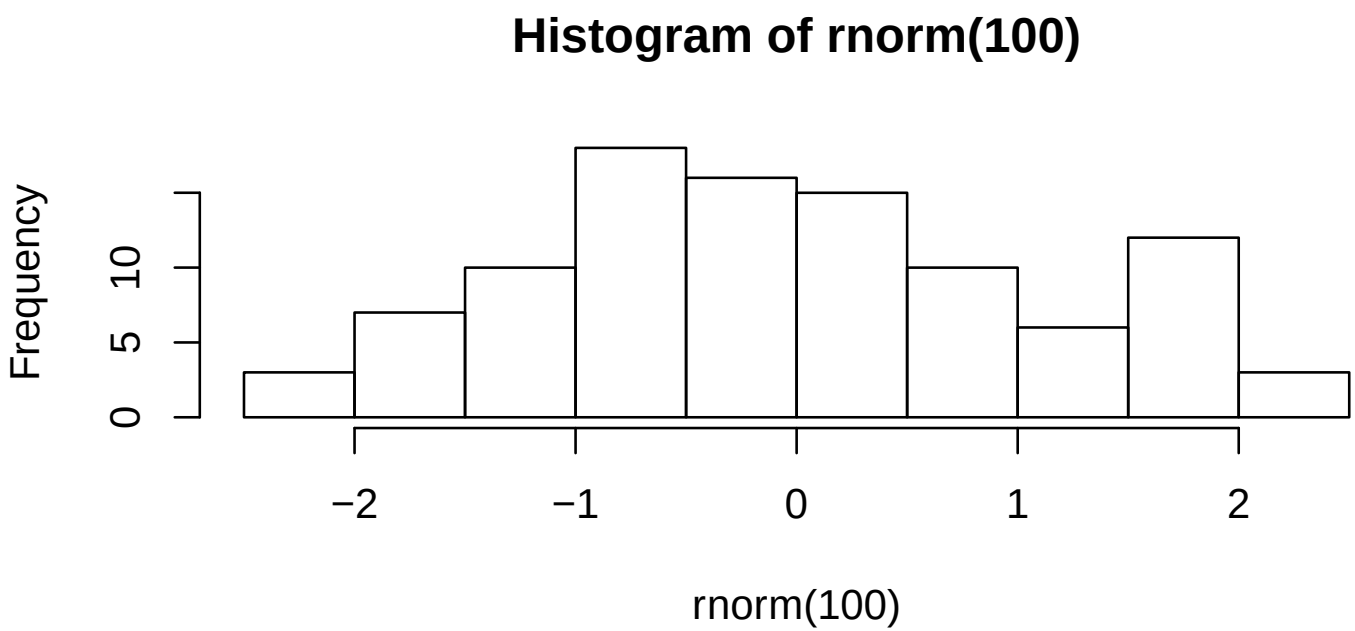
Sweave makes it easy to include figures in your document

```
<<myfig,fig=T,echo=F,include=T,height=3>>=
set.seed(2)
hist(rnorm(100))
@
```

Sweave - figures

Sweave makes it easy to include figures in your document

```
<<myfig,fig=T,echo=F,include=T,height=3>>=
set.seed(2)
hist(rnorm(100))
@
```



Sweave - figures

Sweave makes it easy to include figures in your document

```
<<myfig,fig=T,echo=F,include=T,height=3>>=
set.seed(2)
hist(rnorm(100))
@
```

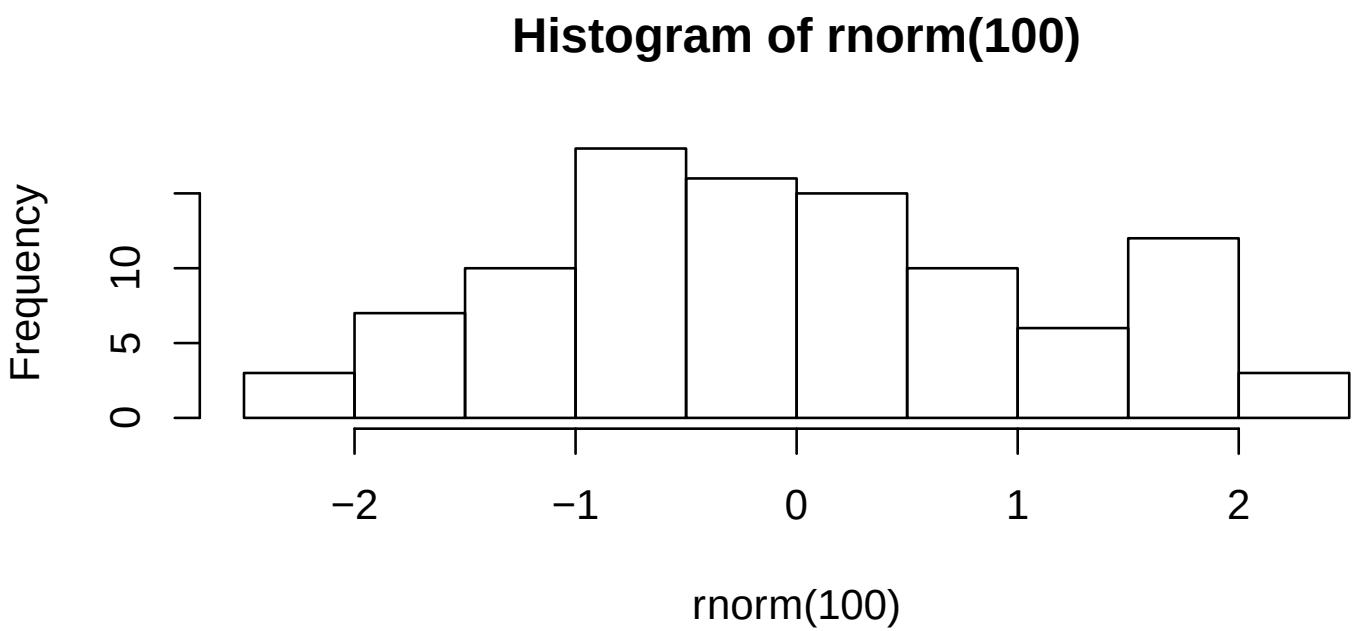
Relevant code options for figures:

- The chunk name is used to name the figure, myfile-myfig.pdf
- `fig`: Lets R know the output is a figure
- `echo`: Use `F` to suppress figure code
- `include`: Should the figure be automatically include in output
- `height`: (and `width`) Set dimensions of figure in inches

Sweave - figures

An alternative approach for creating a figure

```
<<myfig,fig=T,echo=F,include=F,height=3>>=
set.seed(2)
hist(rnorm(100))
@
\includegraphics{rnw_name-myfig.pdf}
```



Sweave - tables

Really easy to create tables

```
<<results=tex,echo=F>>=
library(stargazer)
data(iris)
stargazer(iris,title='Summary statistics for Iris data')
@
```

Sweave - tables

Really easy to create tables

```
<<results=tex,echo=F>>=
library(stargazer)
data(iris)
stargazer(iris,title='Summary statistics for Iris data')
@
```

Table : Summary statistics for Iris data

Statistic	N	Mean	St. Dev.	Min	Max
Sepal.Length	150	5.843	0.828	4.300	7.900
Sepal.Width	150	3.057	0.436	2.000	4.400
Petal.Length	150	3.758	1.765	1.000	6.900
Petal.Width	150	1.199	0.762	0.100	2.500

Sweave - tables

Really easy to create tables

```
<<results=tex,echo=F>>=
library(stargazer)
data(iris)
stargazer(iris,title='Summary statistics for Iris data')
@
```

`results` option should be set to 'tex' (and `echo=F`)

Several packages are available to convert R output to LaTeX table format

- xtable: most general package
- hmisc: similar to xtable but can handle specific R model objects
- stargazer: fairly effortless conversion of R model objects to tables

Sweave - expressions

All objects within a code chunk are saved in the workspace each time a document is compiled (unless `eval=F`)

This allows the information saved in the workspace to be reproduced in the final document as inline text, via expressions

Sweave - expressions

All objects within a code chunk are saved in the workspace each time a document is compiled (unless `eval=F`)

This allows the information saved in the workspace to be reproduced in the final document as inline text, via expressions

```
<<echo=F>>=
data(iris)
dat<-iris
@

Mean sepal length was \Sexpr{mean(dat$Sepal.Length)}.
```

Sweave - expressions

All objects within a code chunk are saved in the workspace each time a document is compiled (unless `eval=F`)

This allows the information saved in the workspace to be reproduced in the final document as inline text, via expressions

```
<<echo=F>>=
data(iris)
dat<-iris
@

Mean sepal length was \Sexpr{mean(dat$Sepal.Length)}.
```

Mean sepal length was 5.84333333333333.

Sweave - expressions

Change the global R options to change the default output

```
<<echo=F>>=
data(iris)
dat<-iris
options(digits=2)
@

Mean sepal length was \Sexpr{format(mean(dat$Sepal.Length))}.
```

Mean sepal length was 5.8.

Sweave vs Knitr

Does not automatically cache R data on compilation

Knitr is a useful alternative - similar to Sweave but with minor differences in args for code chunks, more flexible output

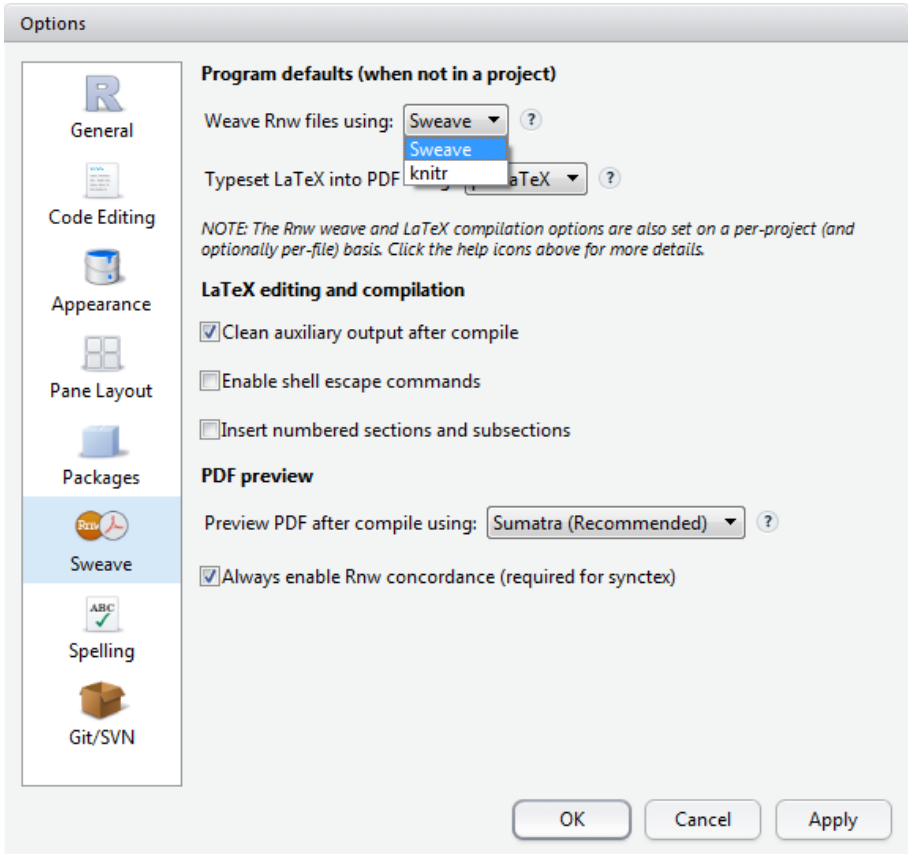
Sweave vs Knitr

Does not automatically cache R data on compilation

Knitr is a useful alternative - similar to Sweave but with minor differences in args for code chunks, more flexible output

Must change default options in RStudio

Knitr included with RStudio, otherwise download as package



Knitr

Knitr can be used to cache code chunks

Date are saved when chunk is first evaluated, skipped on future compilations unless changed

This allows quicker compilation of documents that import lots of data

```
<<mychunk, cache=TRUE, eval=FALSE>>=
load(file='mydata.RData')
@
```

Knitr

.Rnw file

```
\documentclass{article}

<<setup, include=FALSE, cache=FALSE>>=
library(knitr)

#set global chunk options
opts_chunk$set(fig.path='M:/docs/figs/', fig.align='center',
dev='pdf', dev.args=list(family='serif'), fig.pos='!ht')

options(width=60)
@

\begin{document}

Here's some R code:

<<eval=T, echo=T>>=
set.seed(2)
rnorm(10)
@

\end{document}
```

Knitr

The final product:

Here's some R code:

```
set.seed(2)
rnorm(10)

##  [1] -0.89691  0.18485  1.58785 -1.13038 -0.08025  0.13242
##  [7]  0.70795 -0.23970  1.98447 -0.13879
```

Knitr

Figures, tables, and expressions are largely the same as in Sweave

Figures

```
<<myfig,echo=F>>=
set.seed(2)
hist(rnorm(100))
@
```

Tables

```
<<mytable,results='asis',echo=F,message=F>>=
library(stargazer)
data(iris)
stargazer(iris,title='Summary statistics for Iris data')
@
```

A minimal working example

Step by step guide to creating your first RR document

A minimal working example

Step by step guide to creating your first RR document

1. Download and install RStudio

A minimal working example

Step by step guide to creating your first RR document

1. Download and install RStudio
2. Dowload and install MikTeX if using Windows

A minimal working example

Step by step guide to creating your first RR document

1. Download and install RStudio
2. Dowload and install MikTeX if using Windows
3. Create a unique folder for the document - This will be the working directory

A minimal working example

Step by step guide to creating your first RR document

1. Download and install RStudio
2. Dowload and install MikTeX if using Windows
3. Create a unique folder for the document - This will be the working directory
4. Open a new Sweave file in RStudio

A minimal working example

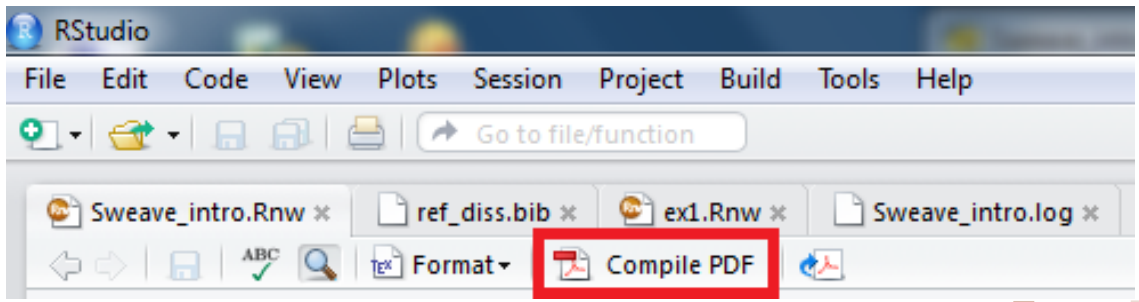
Step by step guide to creating your first RR document

1. Download and install RStudio
2. Dowload and install MikTeX if using Windows
3. Create a unique folder for the document - This will be the working directory
4. Open a new Sweave file in RStudio
5. Copy and paste the file found on slide 7 for Sweave or slide 21 for Knitr into the new file (and select correct compile option)

A minimal working example

Step by step guide to creating your first RR document

1. Download and install RStudio
2. Dowload and install MikTeX if using Windows
3. Create a unique folder for the document - This will be the working directory
4. Open a new Sweave file in RStudio
5. Copy and paste the file found on slide 7 for Sweave or slide 21 for Knitr into the new file (and select correct compile option)
6. Compile the pdf (runs Sweave/Knitr, then pdfLatex)



If things go wrong...

LaTeX Errors can be difficult to narrow down - check the log file

Sweave/Knitr errors will be displayed on the console

Other resources

- 'Reproducible Research with R and RStudio' by C. Garund, CRC Press
- LaTeXforum (like StackOverflow)
  http://www.latex-community.org/forum/
- Comprehensive Knitr guide http://yihui.name/knitr/options
- Sweave user manual
  http://stat.ethz.ch/R-manual/R-devel/library/utils/doc/Sweave.pdf
- Intro to Sweave
  http://www.math.ualberta.ca/ mlewis/links/the˙joy˙of˙sweave˙v1.pdf