Reproducible Research using R

Miqdad Asaria
Centre for Health Economics, University of York

Slides and sample code available at: https://github.com/miqdadasaria/reproducible_research

Irreproducible Research

- There has been a lot of concern recently about scientific results not being reproducible
- Potti-Nevins scandal
  - research unravelled when people tried to reproduce it
  - several published studies retracted
  - 3 major clinical trials cancelled
- Only 6 out of 53 landmark cancer studies could be reproduced – Begley & Ellis, Nature (2012)

Reproducible Research

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures. (Buckheit & Donoho)

Reproducible Research

In this presentation I will use reproducible research to mean a situation where the final research document can be automatically generated from the raw data and code provided alongside it and exactly reproduced by this process.

Scientific reproducibility goes far beyond this simple definition but this helps.

Why Reproducible Research

- When methods, assumptions or data change
- When we want to share our research with others to build on
- When we want to document our analysis as we go along
- When we want to eliminate transcribing errors

Typical Workflow

- Data entry
- Data cleaning and preparation
- Statistical analysis
- Using statistical analysis in modelling
- Tables and figures from model results
- Write up in paper
- Iterate …

Typical Workflow Problems

- Often done in multiple different programs
- Errors often occur at interface – e.g. copy/paste; transcribing; version control
- Time consuming and frustrating

Sweave

- Sweave = R + LaTeX
- R code directly embedded into latex document
- Entire workflow can be captured in the final document ensuring consistency and ease of iteration

LaTeX

- Markup language for writing reports rather than WYSIWYG
- Focus is on content and structure, LaTeX handles layout for you
- Very good for writing equations
- Steep learning curve
- MikTeX is a Windows implementation

Sweave Example 1

We start by creating a basic latex document save with Rnw extension and compile

\documentclass[a4paper]{article}

\begin{document}

\end{document}

Sweave Example 1 continued

We will use the attitude data set in R for our examples so lets introduce it

```
\documentclass[a4paper]{article}
\title{The Chatterjee-Price Attitude Data}
\author{Miqdad Asaria}
\begin{document}

\maketitle

\section{Introduction}
From a survey of …

\end{document}
```

Sweave Example 2

Next we add some inline R code using the \Sexpr{} syntax

\Sexpr{nrow(attitude)}

Sweave Example 3

- We next use some R code to produce LaTeX tables exploring the data
- R code is put in chunks in the Sweave document between <<>>= and @

```
<<echo=FALSE, results=tex>>=
        library(stargazer)
        stargazer(head(attitude), title="Summary of the Chatterjee-Price Attitude Data", summary=TRUE)
@
```

- In this example we use the stargazer R package to produce our tables but there are several alternative such as xtable, apsrtable, memisc, texreg and outreg

Sweave Example 4

We next run some regressions on our data and output the results

```
linear.model.1 = lm(rating ~ complaints + privileges +
learning + raises + critical, data=attitude)
linear.model.2 = lm(rating ~ complaints + privileges +
learning, data=attitude)

attitude$high.rating <- (attitude$rating > 70)
probit.model = glm(high.rating ~ learning + critical + advance,
data=attitude, family = binomial(link = "probit"))
```

Sweave Example 4 continued

We are able to produce publication quality results tables without the need for manual transcribing

```
stargazer(linear.model.1,
linear.model.2,
probit.model,
title="Regression Results",
align=TRUE,
dep.var.labels=c("Overall Rating","High Rating"),
covariate.labels=c("Handling of Complaints","No Special Privileges","Opportunity to Learn","Performance-Based Raises","Too Critical","Advancement"),
float=FALSE)
```

Sweave Example 5

In this next example we add a figure to our report

```
\begin{figure}
\centering
<<echo=FALSE, fig=TRUE>>=
 # R code to create figure
@
\caption{Add an appropriate caption for the figure here}
\end{figure}
```

Sweave Example 5 continued

- We generate a new variable for use in our figure
- As this variable is randomly generated to reproduce our results we must set a seed

```
set.seed(123)
hair_colour = c("Red","Yellow","Black")
attitude$manager_hair_colour =
        factor(hair_colour[round(runif(nrow(attitude))*2) + 1])
```

Sweave Example 5 continued

- We use this new variable to produce the plot

```
library(ggplot2)
ggplot(attitude) +
geom_density(aes(rating, fill=manager_hair_colour), alpha = 0.55) +
scale_fill_manual(name="Manager Hair Colour", limits=hair_colour, values=hair_colour)
```

Sweave Example 6

- For our final example we see what happens if we get some additional data

```
set.seed(123)
# add two new rows into the data set to simulate new responses gathered
new_row_1 = round(runif(ncol(attitude))*100)
new_row_2 = round(runif(ncol(attitude))*100)
attitude = rbind(attitude,new_row_1,new_row_2)
```

Sweave Example 6 continued

- We can see that all our analysis updates seamlessly
- The final step is to display the session information so that others can reproduce our computing environment

```
<<echo=TRUE>>=
    print(sessionInfo(),locale=FALSE)
@
```

Other Useful Tools

- Knitr recently released, very similar to Sweave but much easier to extend to output to other formats e.g. Markdown
- Other alternatives available for other languages e.g. SASweave
- Project code for published studies can be uploaded to http://www.runmycode.org

Limitations

- No track changes for co-authoring documents – need to use version control e.g. GitHub
- Data often confidential might not be able to share it
- Intellectual property in code

Further Reading

- Wavelab and Reproducible Research - Buckheit & Donoho http://statweb.stanford.edu/~wavelab/Wavelab_850/wavelab.pdf
- Department of Biostatistics – Venderbilt University http://biostat.mc.vanderbilt.edu/wiki/Main/SweaveLatex
- Toga ware – One Page R http://togaware.com/onepager/
- Sweave some first steps – Keith Baggerly http://bioinformatics.mdanderson.org/SweaveTalk/sweaveTalkb.pdf
- Knitr http://yihui.name/knitr/
- Sweave http://www.stat.uni-muenchen.de/~leisch/Sweave/
- The (Not So) Short Introduction to LaTeX2e - Tobias Oetiker http://www.ctan.org/tex-archive/info/lshort/english

Further Reading

- Challenges in Irreproducible Research (Nature) http://www.nature.com/nature/focus/reproducibility/index.html
- An array of errors (The Economist) http://www.economist.com/node/21528593
- Trouble at the lab (The Economist) http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble
- Open Science and Reproducible Research (BMJ) http://www.bmj.com/content/344/bmj.e4383