Reproducible Research and R Workflow

Introduction

Reproducible Research and R Workflow Melboure R Users Group (melbURN)

Jeromy Anglim

Psychological Sciences, University of Melbourne

1st December 2010

jeromyanglim.blogspot.com

Outline

Reproducible Research and R Workflow

Introduction

- 1 Introduction
- 2 Workflow
- Tools
 - Eclipse and StatET
 - Version Control
 - make and makefiles
 - Sweave and LaTeX
- 4 Sweave Examples
 - 1. Console Report
 - 2. Multiple Reports
 - 3. Database Driven Document.
 - 4. Non-console Report

Quote from John Chanmbers

Reproducible Research and R Workflow

> Jeromy Anglim

Introduction

147 171

Lool

Eclipse and StatET Version Control make and makefiles Sweave and LaTeX

Sweave Examples

Reports

Driven Document

4. Non-conso

The Mission:

Enable the best and most thorough exploration of data possible.

..

The Prime Directive:

The computations and the software for data analysis should be trustworthy.

Source: John M. Chambers, Chapter 1, Software For Data Analysis: Programming with R

What is the End Product?

Reproducible Research and R Workflow

> Jeromy Anglim

Introduction

Workflow

Too

StatET
Version Control
make and
makefiles
Sweave and

Sweave Example

Console
Report
 Multiple

 Database Driven Document
 Non-consol Report

- Console displayed versus no console displayed
- Batch versus once off
- Data:
 - Cleaned
 - Processed
 - Documented
- Data anlysis software:
 - R Package
 - A model

Focus of this talk

A workflow for writing reproducible data driven reports

The Initial Challenge for the R Learner

Reproducible Research and R Workflow

Introduction

How should you

- divide a project into files and folders?
- incorporate R analyses into a report?
- convert default R output into publication quality tables, figures, and text?
- build the final product?
- sequence the analyses?
- divide code into functions?

i.e., How do you efficiently achieve the Mission and fulfill the Prime Directive?

David Smith's Tips on R Workflow

Reproducible Research and R Workflow

> Jeromy Anglim

Introducti

Workflow

Tool

StatET
Version Control
make and
makefiles
Sweave and
LaTeX

Example

Report

2. Multiple

Driven Document Transparency: Logical organisation of units

Maintanability: Standardisation, clear comments

Modularity: DRY Principle, Discrete units

 Portability: Relative paths, minimise dependencies, dependencies are clear

Reproducibility: Easy to reproduce results

Efficiency: Easy to maintain and modify

Source: http://blog.revolutionanalytics.com/2010/10/a-workflow-for-r.html

Josh Reisch LCFD Model

Reproducible Research and R Workflow

Workflow

■ load.R

2 clean.R

3 func.R

4 do.R

Source: http://stackoverflow.com/questions/1429907/

workflow-for-statistical-analysis-and-report-writing/1434424

John Myles White and ProjectTemplate

Reproducible Research and R Workflow

> Jeromy Anglim

Introduct

Workflow

TTOTALIO

Eclipse and StatET Version Contro make and

make and makefiles Sweave and LaTeX

Sweave Examples

Report

2. Multiple

 Multiple Reports
 Database

3. Database Driven Document 4. Non-conso

Best practice ideas

- Efficient creation of new projects
- Standardised folder and file structure (i.e., data, diagnostics, doc, graphs, lib, logs, profiling, reports, tests)
- Automatic data loading
- README and TODO files
- Encourages unit testing
- Standardised location of library() statements
- and more ...

ProjectTemplate

```
Reproducible
Research and
R Workflow
```

Jeromy Anglim

Introduct

Workflow

Тоо

StatET
Version Control
make and
makefiles
Sweave and
LaTeX

Example Example

2. Multiple Reports 3. Database

Driven Document 4. Non-consol

```
install.packages('ProjectTemplate')
library('ProjectTemplate')
?ProjectTemplate

create.project('my-project')

setwd('my-project')

load.project()
```

See also http://www.johnmyleswhite.com/notebook/2010/08/26/projecttemplate/

R Programming Environments

Reproducible Research and R Workflow

> Jeromy Anglim

.....

Workflow

_ .

Eclipse and StatET

Version Control make and makefiles Sweave and

Examples
1. Console

2. Multiple Reports

Driven
Document

4. Non-consol

- Rgui
- Emacs + ESS
- Eclipse + StatET
- Any text editor + command line
- and many more . . .

Eclipse and StatET: Screenshot

StatETVerticalSplit - Sweave_MCQ/Sweave_MCQ.Rnw - Eclipse

Reproducible Research and R Workflow

> Jeromy Anglim

Introduction

Workflow

lools

Eclipse and StatET

Version Contro make and makefiles Sweave and LaTeX

Sweave Example

Report

2. Multipl

2. Multip Reports 3. Databa

Document
4. Non-const

File Edit Source Navigate Search Project Run Document Latex Symbols Window Help TP ar StatETVerticalSole air Markdown air <StatET> In Resource air StatETLaptop P 23 ≥ 0 3 " T *RMeetup Workflow.te makefile Sweave MCO.Rnw E3 " R IR Consolel R-2.11.1 / RJ (01/12/2010 4:05:58 PM) - C:\Users\ ... \2010 Sweave Examples\Sweave M 1\documentclass[12pt, aipaper](exam) 2\usepackage(GT1)(fontenc) > makefile x - a x & | 0, a | e 0 - e -3\usepackage(Sweave) Rmeetup Workflow 1 4\SweaveOpts(echo=FALSE) Rmeetup Workflow I Welcome at Wed Dec 01 16:05:59 2010 S\usepackage(hyperref) % for links (only use on final pri A > RMeetup Workflov 6\hvpersetup(pdfpagelavout=SinglePage) % http://www.tug.org/applicat# > items <- read.csv("data/items.csv", stringsAsFactors = FALSE) M TODO,md 7\setkevs(Gin)(width=0.8\textwidth) Sweeze Item Analysis (S-8\pagestyle(headandfoot) % every page has a header and footer Swence MCO (Swence I) 9\header()(Sample Multiple Choice Questions)() a backup 10\footer()(Page \thepage\ of \numpages)() @ .output G data 12\begin(document) agitignore. 13<<pre>prepare data>>= nmiect 14items <- read.csv("data/items.csv", stringsAsFactors = FALSE) makefile README.md 16writeQuestion <- function(x) (Sweave_MCQ.Rnw 17 o("\\filbreak", paste("\\question\n", x["itemText"]), Sweave.sty "\\begin(choices)", Sweave_Personality_Repo paste("\\choice", x("optionA")), 24 > Winter_Olympics [Wi paste("\\choice", x["optionB"]), paste("\\choice", x["optionC"]), @ Cmd (\$ Queu 23 paste("\\choice", x["optionD"]), R [R Console] "\\vspace(10 mm)" EL <idle> Search 33 'git' - 158 matches in 'Blog' (".") iii Blog a drafts Academic_Projects_and_Repositories.md Additional Tips Sweave Eclipse.md bloggingWorkflow.html (3 matches) 4 11 Smart Insert 15:1

Eclipse and StatET: Benefits

Reproducible Research and R Workflow

> Jeromy Anglim

Introduc

Workflox

Eclipse and

StatET
Version Contro
make and
makefiles
Sweave and
LaTeX

Examples 1. Consol

Report

Reports 3. Database Driven

Driven
Document
4. Non-console
Report

Good support for version control

Easy to hook in external tools like sh, cmd, and make

■ File search

Allows for multiple integrated consoles

 Configurable multi-element display (particularly good on big monitors)

 Understands R (indentation, colour coding, code folding, outline view)

 Great shortcut keys for sending R code to console and getting help

Understands Sweave and LaTeX

Project explorer for projects, folders, files

R object explorer and content assist

Command history and Queue

Eclipse and StatET: Resources

Reproducible Research and R Workflow

> Jeromy Anglim

Workflow

Eclipse and

Version Control make and makefiles Sweave and LaTeX

Example

Report

2. Multiple

Reports

4. Non-conso

StatET Website:

http://www.walware.de/goto/statet

■ Longhow Lam's Guide:

http://www.splusbook.com/RIntro/RCourseMaterial.html

My Guide:

Version Control: Practical Benefits

Reproducible Research and R Workflow

> Jeromy Anglim

.....

Workflow

- ..

Version Control make and makefiles

make and makefiles Sweave and LaTeX

Example

2. Multiple Reports

3. Datab Driven Documer

4. Non-consol

 Rewind a project or a file to a previous state (encourages experimentation)

- Provides a record of changes
- Facilitates collaboration
- Facilitates backup
- Shows changes between files
- Facilitates code sharing and reproducibility

Version Control: Conceptual Benefits

Reproducible Research and R Workflow

> Jeromy Anglim

Introductio

Workflow

lool

Version Control
make and
makefiles
Sweave and
LaTeX

Sweave Example

Report

2. Multiple

Driven Document

4. Non-conso

- the distinction between source and derived files
- the nature of dependencies:
 - dependencies between elements of code
 - dependencies between files within a project
 - and dependencies with files and programs external to the repository
- the nature of a repository and how repositories should be divided
- the nature of committing and documenting changes and project milestones

Git: A Version Control System

Reproducible Research and R Workflow

> Jeromy Anglim

Introductio

....

Workflow

E-0---

StatET

Version Control make and makefiles Sweave and

Sweave Examples

Console
Report

2. Multiple Reports

Driven Document

4. Non-consol

- Popular
- Github
- Experts (e.g., Handley Wickham, Linus Torvalds)

EGit: A Git plugin for Eclipse

Reproducible Research and R Workflow

Version Control

Simple graphical interface integrated with Eclpise

Good for getting started with version control

Tutorial on Getting Started:

http://jeromyanglim.blogspot.com/2010/11/getting-started-with-git-egit-eclipse.html

make and makefiles

Reproducible Research and R Workflow

> Jeromy Anglim

Introduct

Workfloy

Eclipse and StatET Version Control make and makefiles Sweave and

Sweave Example

Report

2. Multiple

Driven
Document

4. Non-conso

One-click build

Efficient build

Reliable build

Separate source from derived files

Clean derived files

Run alternative builds

Encourages clear thinking about dependencies

Tutorial on getting started:

Example makefile

```
Reproducible
Research and
R Workflow
            output = .output
            rnwfile = Sweave_MCQ
            backup = .backup
            all:
                R CMD Sweave $(rnwfile).Rnw
                 -mkdir $(output)
                -cp *.sty $(output)
make and
                -mv *.tex *.pdf *.eps $(output)
makefiles
                cd $(output); texify --run-viewer --pdf $(rnwfile).tex
            clean.
                -rm $(output)/*
            backup:
                -mkdir $(backup)
                cp $(output)/$(rnwfile).pdf $(backup)/$(rnwfile).pdf
```

Sweave

Reproducible Research and R Workflow

> Jeromy Anglim

minoduci

Workflo

Eclipse and StatET Version Control make and makefiles Sweave and LaTeX

Example

2. Multiple Reports 3. Database

Driven Document 4. Non-consol Weave S (i.e., R) code chunks with LaTeX in a single self-describing document.

Key Benfits

- Reproducibility
- Efficiency
- Reliability
- Education & Communication
- Manual: http://www.stat.uni-muenchen.de/~leisch/Sweave/
- My guide to getting started: http://jeromyanglim.blogspot.com/2010/02/getting-started-with-sweave-r-latex.html

Overview of Examples

Reproducible Research and R Workflow

> Jeromy Anglim

IIItroduct

Workflov

Tool

Eclipse and StatET Version Control make and makefiles Sweave and LaTeX

Sweave Examples

Report

2. Multiple

Driven
Documen

4. Non-consol

Different Types of Sweave Documents

- Console Report
- Multiple Reports
- Database Driven Document
- Non-console Report

For each example links are provided to complete copies of source code with explanation.

1. Console Report: Item Analysis

Reproducible Research and R Workflow

> Jeromy Anglim

Introduction

M/------

Tools

Eclipse and StatET Version Contro make and makefiles Sweave and

Sweave

 Console Report

2. Multiple

3. Database Driven

4. Non-consol

http://jeromyanglim.blogspot.com/2010/11/sweave-tutorial-3-console-input-and.html

2. Multiple Reports: Personality Feedback

Reproducible Research and R Workflow

2. Multiple Reports

http://jeromyanglim.blogspot.com/2010/11/sweave-tutorial-2-individual.html

3. Database Driven Document: Multiple Choice Questions

Reproducible Research and R Workflow

> Jeromy Anglim

Introduction

147 171

Tool

Eclipse and StatET Version Contro make and makefiles Sweave and LaTeX

Sweave Example

1. Console

Reports
3. Database
Driven

Document
4. Non-consc

http:

4. Non-console Report: Winter Olympic Medals

Reproducible Research and R Workflow

> Jeromy Anglim

Introduction

147 171

.....

1001

Eclipse and StatET Version Contro make and makefiles Sweave and LaTeX

Sweave Examples

1. Console Report

2. Multiple Reports

3. Database Driven

4. Non-console Report https://github.com/jeromyanglim/Sweave_Winter_Olympics