Reproducible Research Notes

Coursera Course by John Hopkins University

INSTRUCTORS: Dr. Jeff Leek, Dr. Roger D. Peng, Dr. Brian Caffo

Contents

Intro

- Reproducible Research applies to data analysis but also any sort of processing of data to help convey what has been done to the data so an analyis can be reproduced in the future.

- This course will cover the tools one can use in R to communicate what one has done with the data

Course Description

- "**In this course you will learn the ideas of reproducible research and reporting of statistical analyses. Topics covered include literate programming tools, evidence-based data analysis, and organizing data analyses. In this course you will learn to write a document using R markdown, integrate live R code into a literate statistical program, compile R markdown documents using knitr and related tools, publish reproducible documents to the web, and organize a data analysis so that it is reproducible and accessible to others.**"

Course Book

(The book can be downloaded as a pdf from leanpub)[https://leanpub.com/reportwriting]

What is Reproduciblity about?

- Peng makes an analogy between data science and music, he compares two songs: + **(Code Monkey)[https://www.youtube.com/watch?v=qYodWEKCuGg]**
  + **(Symphony No. 8)[https://www.youtube.com/watch?v=e7WgXhUBrps]**
- The second song is quite complex, it's even been nicknamed "Symphony of a Thousand" for the amount of people required to perform it. The score that comes with it gives detailed information of what every section is to be doing during the piece.
- In addition, *Mahler* was a conducter and often felt frustrated with scores that had complex parts but didn't convey enough information about what the composer wanted. So when he wrote his music he wrote detailed instructions with the score.
- In Data analysis there is no one unified way that the "score" of a data analysis is conveyed. As such everyone has thier own way from describing what was done to providing all the code. The first can sometimes be lacking and the second can seem to be an information overload.

Concepts, Ideas, & Structure

Concepts and Ideas (Part 1)

Replication

- The ulitimate standard for strengthening scientific evidence is replication of findings and conducting studies with independent: + Investigators
  + Data
  + Analytical methods
  + Laboratories
  + Instruments

- Replication is particularly important in studies that can impact broad policy or regulatory decisions

However, * Some studies cannot/can be challenging to be replicated
+ No time, studies nowadays require large sample sizes
+ No money, researchers gotta eat too
+ Unique, sometimes a study is of a particular subset (Air Pollution, 'rona)
* Reproducible Research makes analytic data and code available so that others may reproduce findgs; a middle ground between replication and nothing

Why Do We Need Reproducible Research?

- New technologies increasing data collection throughput; data are more complex and extremely high dimensional

- Existing data bases can be merged into new "megadatabases"

- Computing power is greatly increased, allowing more sophisticated analyses
  + Kinda like using DNA evidence for old cold cases

- For every field "X" there is a field "Computational X"
  + Reproducing the Computational X from the X will allows others to be confident the correct analysis was done

Example: Reproducible Air Pollution and Health Research

- Estimating small (but important) health effects in the presence of much stronger signals
  + Air pollution lightly impacts health but still effects it enough.. on occasion

- Results inform substantial policy decisions, affect many stakeholders
  + EPA regulations can cost billions of dollars, so the research must be reproducible to convey the reason for the need of these new regulations

- Complex statistical methods are needed and subjected to intense scrutiny

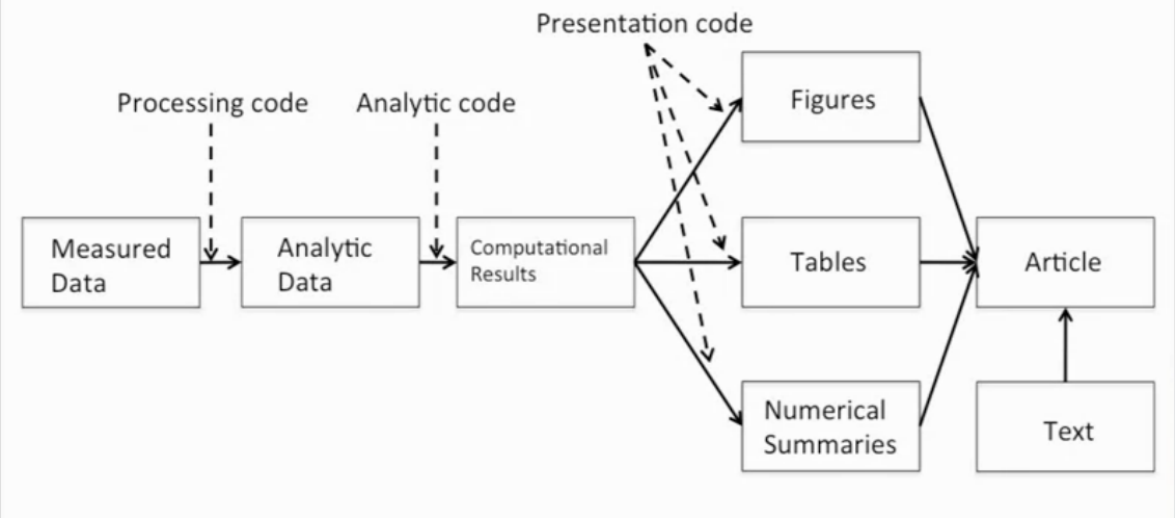
- See Also: Internet-based Health and Air Pollution Surveillance System (iHAPSS)

Concepts and Ideas (Part 2)

Research Pipeline

- When you read an article you only get the article, not the data that are behind it.

- This is where the research pipeline comes in..



Recent Developments in Reproducible Research

(The Duke Saga)[https://www.youtube.com/watch?v=eV9dcAGaVU8&feature=emb_err_watch_on_yt]
(Evolution of Translational Omics: Lessons Learned and the Path Forward)[https://www.nap.edu/catalog/13297/evolution-of-translational-omics-lessons-learned-and-the-path-forward]
* In the Discovery/Test Validation stage of omics-based tests:
+ **Data/metadata** used to develop test should be made publicly available
+ The **computer code** and fully specified computational procedures used for development of the candidate omics-based test should be made sustainably available
+ "Ideally, the computer code that is released will **encompass all of the steps of computational analysis**, including all data preprocessing steps, that have been described in this chapter. All aspects of the analysis need to be transparently reported."

What do We Need for Reproducible Research?

- Analytic data are available
- Analytic code are available
- Documentation of code and data
- Standard means of distribution

Who are the Players: * Authors
+ Want to make their research reproducible
+ Want tools for RR to make their lives easier (or at least not much harder)
* Readers
+ Want to reproduce (and perhaps expand upon) interesting findings
+ Want tools for RR to make their lives easier

Challenges

- Authors must undertake considerable effor to put data/results on the web (may not have resources like a web server)
- Readers must download data/results individually and piece together which data go with which code sections, etc.
- Readers may not have the same resources as authors
- Few tools to help authors/readers (although toolbox is growing!)

In Reality...
* Authors + Just put stuff on the web
+ (Infamous) Journal supplementary materials skewed about

+ There are some central databases for various fields (e.g. biology, ICPSR)
* Readers
+ Just download the data and (try to) figure it out
+ Piece together the software and run it

Concepts and Ideas (Part 3)

Literate (Statistical) Programming

- An article is a stream of **text** and **code**
- Analysis code si divided into text and code "chunks"
- Each code chunk loads data and computes results
- Presentation code formats results (tables, figures, etc.)
- Article text explains what is going on
- Literate programs can be **weaved** to produce human-readable documents and **tangled** to produce machine-readable documents
- Literate programming is a general concept that requires:

1) A documentation language (human readable)
2) A programing language (machine readable)

Sweave

- Pronounced S-weave
- Uses L[A]T_E_X (Pretend that worked) and R as the documentation and programming languages
- Sweave was developed by Friedrich Leisch (member of the R Core) and is maintained by R core
- **Website**

Limitations: * Focused primarily on LaTeX, a difficult to learn markup language used "only by weirdos"
* Lacks features like caching, multiple plots per chunk, mixing programming languages and many other technical items
* Not frequently updated or very actively developed

knitr

- knitr si an alternative (more recent) package
- Brings together many features added on to Sweave to address limitations
- knitr uses R as the programming language (although others are allowed) and variety of documentation languages
  + LaTeX, Markdown, HTML
- knitr was developed by Yihui Xie (while a graduate student in statistics at Iowa State)
- **Website**

Reminder to commit, delete this line *AFTER* committing

Summary

- Reproducible research is important as a **minimum standard**, particularly for studies that are difficult to replicate
- Infrastructure is needed for **creating** and **distributing** reproducible documents, beyhond what is currently available
- There is a growing number of tools for creating reproducible documents

Scipting Your Analysis

Structure of a Data Analysis (Part 1)

Structure of a Data Analysis (Part 2)

Organizing Your Analysis

Reminder to commit, delete this line *AFTER* committing

Quiz 1

Reminder to commit, delete this line *AFTER* committing

Markdown & knitr

Coding Standards in R

Markdown

R Markdown

R Markdown Demo

Reminder to commit, delete this line *AFTER* committing

knitr (Part 1)

knitr (Part 2)

knitr (Part 3)

knitr (Part 4)

Reminder to commit, delete this line *AFTER* committing

Quiz 2

Reminder to commit, delete this line *AFTER* committing

Course Project 1

Reminder to commit, delete this line *AFTER* committing

Reproducible Research Checklist & Evidence-based Data Analysis

Communicating Results

RPubs

Reminder to commit, delete this line *AFTER* committing

Reproducible Research Checklist (Part 1)

Reproducible Research Checklist (Part 2)

Reproducible Research Checklist (Part 3)

Reminder to commit, delete this line *AFTER* committing

Evidence-based Data Analysis (Part 1)

Evidence-based Data Analysis (Part 2)

Evidence-based Data Analysis (Part 3)

Evidence-based Data Analysis (Part 4)

Evidence-based Data Analysis (Part 5)

Reminder to commit, delete this line *AFTER* committing

Case Studies & Commentaries

Caching Computations

Case Study: Air Pollution

Reminder to commit, delete this line *AFTER* committing

Case Study: High Throughput Biology

Commentaries on Data Analysis

Reminder to commit, delete this line *AFTER* committing

Course Project 2

Reminder to commit, delete this line *BEFORE* committing