Reproducible Research

Therri Usher

Tuesday, February 3, 2015

Announcements

Office Hours

- Office hours will be held on Wednesdays from 11 AM to 12 PM.
- We are still working on finding a more convenient location for office hours.
- **For this week**, office hours will be held in the Public Health Studies building, room 209.
- If you need to meet with me but cannot attend office hours, please schedule an appointment with me beforehand.

Examples From Last Class

1. Download the test.dta file from Blackboard. Load the dta file into R and assign it to the variable name test.data.
2. Create a sequence of numbers from 1 to 100 and assign it to the variable name x. Save the variable in a .Rdata file to your desktop.

Examples From Last Class - Answers

1. Download the test.dta file from Blackboard. Load the dta file into R and assign it to the variable name test.data.

Answer:

```
install.packages("foreign", dependencies=TRUE)
library("foreign")
test.data <- read.dta("test.dta")
```

2. Create a sequence of numbers from 1 to 100 and assign it to the variable name x. Save the variable in a .Rdata file to your desktop.

Answer:

```
setwd("C:/Users/owner/Desktop")
x <- 1:100
save(x, file="seq.Rdata")
```

What Is Reproducible Research?

"The idea that the ultimate product of academic research 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." - Wikipedia

"The idea that data analyses, and more generally, scientific claims, are published with their data and software code so that others may verify the findings and build upon them." - Coursera

"Storing your data and documenting your analysis in code such that others can recreate the same results." - Me

How Does It Differ From Replication?

Replication is the ability to repeat an entire experiment and obtain the same results.

How Does It Differ From Replication?

Replication is "the standard by which scientific claims are evaluated." Reproducible research is a "candidate" for the "minimum standard that can fill the void between full replication and nothing." - Roger Peng, "Reproducible Research and *Biostatistics*"

Basic Ideas of Reproducible Research

- Code everything.
  - Do not use the interface unless it provides code for what it does.
- Document everything.
  - Comment your code.
  - Indicate packages needed to run the code.
- Obtain data in the rawest form possible.
  - If it is not completely raw, keep track of what changes have been done.
  - Do not go into a data file and start changing things!
- If you choose to create a document, create a live document.
  - No copying and pasting tables, figures, and data.

Thoughts To Help Guide Reproducibility

- Is my data accessible for others?
- Does the code do what I think it does?
- Can others understand my code and what it does?
- Can the data and code reproduce my tables and figures?
- Is the workflow and thought process of the analysis documented?
- Can others reproduce this analysis on their own computers given the data and code?
- Extra: Can others extend this analysis to other areas?

Discussion Time

In your groups, discuss the advantages and disadvantages of reproducible research. Write them on the whiteboard and be prepared to discuss some of them.

Advantages and Disadvantages of Reproducible Research

Advantages

- It helps to avoid mistakes
- Analysis is more likely to be correct
- It saves time in the long run
- Analyses typically have a higher impact

Disadvantages

- It is not easy, particularly for large projects
- It is not always possible to make data available
- It is very often impossible to obtain data in raw form
- It is time consuming in the short term
- It can actually be expensive

Tools for Reproducible Research

- R Markdown
- knitr
- GitHub
- R packages

R Markdown

Markdown allows plain text formatting to be converted to different formats, such as HTML and PDF.

R Markdown is designed to allow for embedded R code into markdown documents in order to create reports from R.

R Markdown versus LaTeX

LaTeXis a typesetting system used for creating technical papers. One disadvantage is that documents created using LaTeXare not living documents.

R Markdown is capable of reading and compiling LaTeXformatting but the document is still considered living.

Example: LaTeX versus LaTeX

You can get tips on using R Markdown by going to Help, Markdown Quick Reference.

knitr

knitr is an “engine for dynamic reprt generation with R.” It enables the integration of R code into different documents, such as LaTeX. - Wikipedia

Growing in popularity, many people use knitr to make live documents.

To use knitr in RStudio, go to Tools, Global Options, click on Sweave, then change “Weave Rnw files using” to knitr.

Active Exercise

Create a R Markdown HTML document. It can be a document or presentation. The document should have at least:

- One header
- One word in bold or italic
- 2-item list
- Block of R code that returns output
- Bonus: One linked phrase

The block of R code can be as simple as a histogram or scatterplot. Be sure to know how to hide the code and the output.

GitHub

GitHub is a repository system that utilizes version control and source code management.

It is not specifically required for reproducible research but it can help, especially for collaborative projects.

Ex: This Class' Repository

Above all, GitHub is a social network.

My Expectations of Reproducible Research For The Project

Required:

- Clear, correct, and commented code

Recommended:

- Available data
- Using R Markdown to create your report

Next Class: Forming A Research Question

- Start thinking about the research question you want to answer in your data analysis project.
- Keep an eye on Blackboard for any resources related to research questions. Feel free to share if you find any!