Teaching data analysis in R through the lens of reproducibility

Mine Çetinkaya-Rundel

Duke University, Department of Statistical Science

Andrew Bray

Mt. Holyoke College, Department of Mathematics and Statistics

Objective

Instill the importance and practice of reproducibility from day one of learning to do data analysis.

Toolkit

Analyze data and generate reproducible reports using R / RStudio with R Markdown and knitr.

Traditional Approach

A Better Approach

find out more

Baumer, B.; Çetinkaya-Rundel, M.; Bray, A.; Loi, L.; & Horton, N.J.(2014). R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics. TISE, 8(1).
http://bit.ly/user2014_reproduce

R Markdown in Intro. Stat.

We have been using R Markdown in introductory statistics courses at **Duke University** and **Smith College** since Fall 2012 (large, non-calc based intro course at Duke and intro course with calc pre-calc + regression course at Smith). At both institutions the **students typically have no programming background**. R Markdown is used for weekly **data analysis labs** as well as individual and team based data analysis **projects**.

Data Analysis Labs

Students are given a template with each lab (to ensure consistency in formatting), and input:

```
#### Exercise 7:
It appears as though the majority of respondents' desired weights are slightly below their current weights. As weight increases, desired weight generally increases as well, so the relationship is relatively strong (positive association, linear). There are a couple of outliers that seem to be inaccurate reportings (ie, desired weight of 600 and 700 lbs.), and it would be wise to disregard these data.
```{r fig.width=8, fig.height=4}
plot(cdc$wtdesired ~ cdc$weight, main = "Weight vs. desired weight")
```
```

and get the following as output:

```
## Exercise 7:
It appears as though the majority of respondents' desired weights are slightly below their current weights. As weight increases, desired weight generally increases as well, so the relationship is relatively strong (positive association, linear). There are a couple of outliers that seem to be inaccurate reportings (ie, desired weight of 600 and 700 lbs.), and it would be wise to disregard these data.

plot(cdc$wtdesired ~ cdc$weight, main = "Weight vs. desired weight")
```

- Interpretations, code, output / plots in one place.
- Much tidier than working in the console and much more complete than an R script.
- Compiling the document runs all code from the beginning, which helps avoid common novice errors.

comments / questions

@minebocek
#user2014

mine@stat.duke.edu

Survey Results

Pedagogical Outcomes

■ Instills early the idea that all analysis must be done in a reproducible framework:

Providing students (scientists-in-training) with the tools to do and present reproducible data analysis before teaching them rigorous methods highlights the importance of reproducibility.

■ Emphasizes iteration:

Establishes the norm that data analysis is an iterative process involving critical feedback and revision (just like in a writing course, where multiple drafts pave the way to a final draft), and inherently makes iteration easy.

■ Eases collaboration:

A single file containing every aspect of the analysis (code, output, plots, text) makes collaboration easier, which can inspire instructors to plan more activities that require collaboration between the students.

■ Removes ambiguity from grading:

Teaching data analysis in the reproducible paradigm means being able to provide valuable and targeted feedback. In addition, grading a markdown document is easier than grading an irreproducible document since all components are contained in a single file and guesswork in inferring how the students reached their conclusions is removed.

Quotes from Students

"Using R Markdown has been so great because we do not spend as much time solving the format/design of the paper and instead focus on actual problem solving."

"There is something satisfying about hitting 'knit' and having the text turn into figures and tables."

"I like it a lot actually. It has sparked an interest in coding for me."