R MARKDOWN, KNITR AND REPRODUCIBLE DOCUMENTATION

Yichen Qin

BANA7038

University of Cincinnati

Why is reproducing results so hard?

- Data, code, description come from different sources.
- Manually piece them together.
- Which goes with which?
- No single document that integrates data, code, and description.

We hope

- One single file that streamlines the data, code, text and results (figures and tables).
- That single file explains everything.
- That single file contains: text chunk and code chunk.
- That one single file is compiled to a human readable file.

How to make my work reproducible?

- Do it at the beginning.
- Keep track of things, e.g. github.
- Software whose operations can be coded.
- Do not save output.

Pros and Cons

- Pros:
  - Everything is in one place.
  - Results are automatically updated.
- Cons:
  - Everything is in one place.
  - Difficult to read if it is long.
  - Slow if it is too long

Knitr, Markdown, and R Markdown

- Knitr is an R package (written by Yihui Xie), availabel on CRAN
- Markdown is a simple version of “markup” language.
  - Easy to read instead of html or latex (markup languages).
- R Markdown is an R version of Markdown.

knitr

- R Markdown → html, pdf, doc

- Built in Rstudio

What is knitr good for?

- Manuals
- Short/medium length documents
- Tutorials
- Periodically generated reports (analytics)
- Data preprocessing documents

What is knitr NOT good for?

- Looooooooooong research articles.
- Time consuming computations.
- Documents require precise formatting.

Template file

- See file: eg.Rmd

Code chunk

~~~
```{r codechunkname, echo=TRUE/FALSE, results=“asis”/“hide”, fig.height=123,fig.width=123}
x=runif(100)
epsilon=rnorm(100)*0.1
y=2+3*x+epsilon
plot(x,y)
```
~~~

Inline code

- `r model1$coef[2]`
- `model1$coef[2]`

Header

- `# Header 1`
- `## Header 2`
- `### Header 3`
- `#### Header 4`
- `##### Header 5`
- `###### Header 6`

Dash

- endash: --
- emdash: ---
- ellipsis: ...

Formatting

- `*italic*`
- `**bold**`
- `superscript^2^`

Table

- A | B | C
- --- | --- | ---
- 1 | Maale | Blue
- 2 | Female | Pink

Insert figure

- `![aaaaaaaaaaaAAAAAAAAAAAA](fig1.png)`

Formulas

- `$y_i = \beta_0 + \beta_0 x_i + e_i$`
- `$$\frac{1}{1+\exp(-x)}$$`

Comment

- `<!-- This is comment -->`

Quotes

- > To be, or not to be, that is the question:
- > Whether 'tis nobler in the mind to suffer
- > The slings and arrows of outrageous fortune,

Cache

- What if one code chunk take a long time to run?
- When you re-knitr the document, the code chunk is re-computed.   Not good.
- **`cache=TRUE`** stores the results from each code chunk.
- You have to name each code chunk.
- After the first run, the results are stored for later use.

Cache

- If the data, code changes, everything is re-computed.
- Dependencies are not checked.
- Use it carefully!