Introduction to R for Epidemiologists

Jenna Krall, PhD

Thursday, April 2, 2015

Reproducible research

What is reproducible research?

"...the calculation of quantitative scientific results by independent scientists using the original datasets and methods"

What do we need?

- Reproducibility = Tools + Workflow (Stodden, Leisch, and Peng)
- Literate statistical programming (Rossini via Knuth):
  - programming language + documentation language
- We need to link **data, code, results, and interpretation**

Introduction to knitr

knitr

- Developed by Yihui Xie at Iowa State (now at RStudio)
- Integration with RStudio

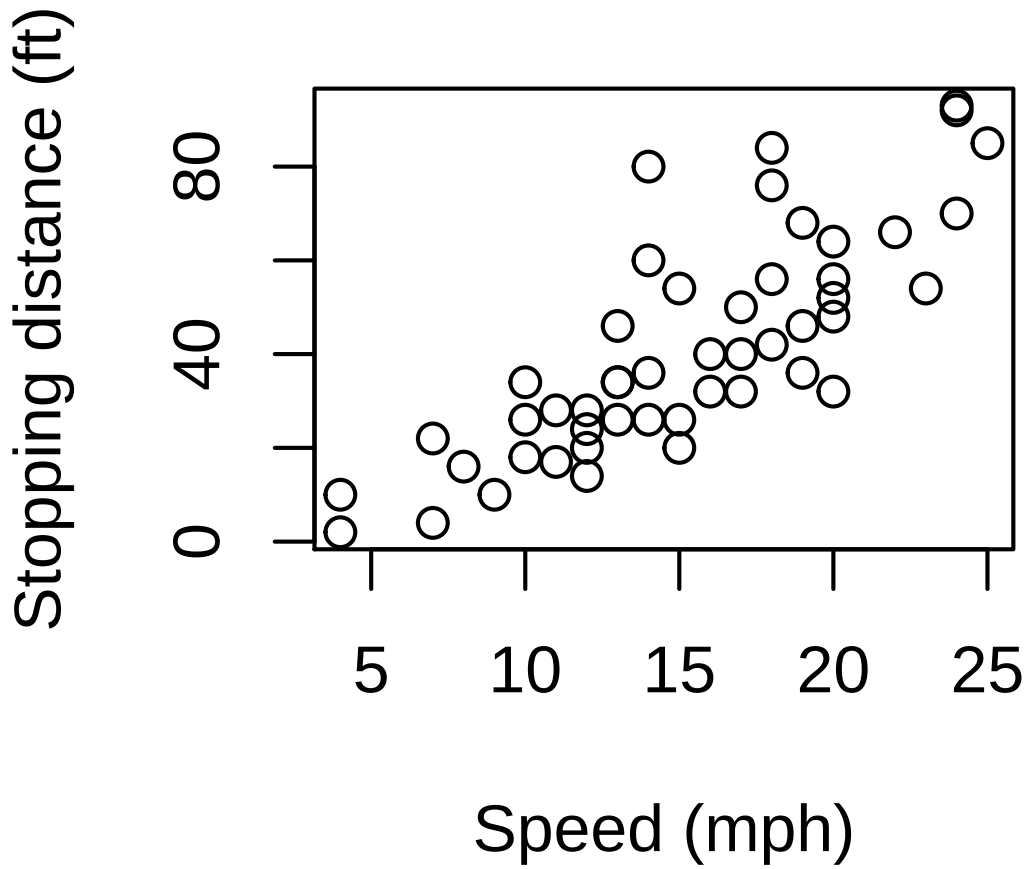
We will focus on using knitr to integrate R and markdown, but knitr is actually a flexible tool that can be used in broader ways.

- We write text (in markdown) separated by code chunks (written in R)
- In our output, we include code (or not) and results (relevant tables and figures)
- knitr ties together our code and our final report

This class is reproducible!

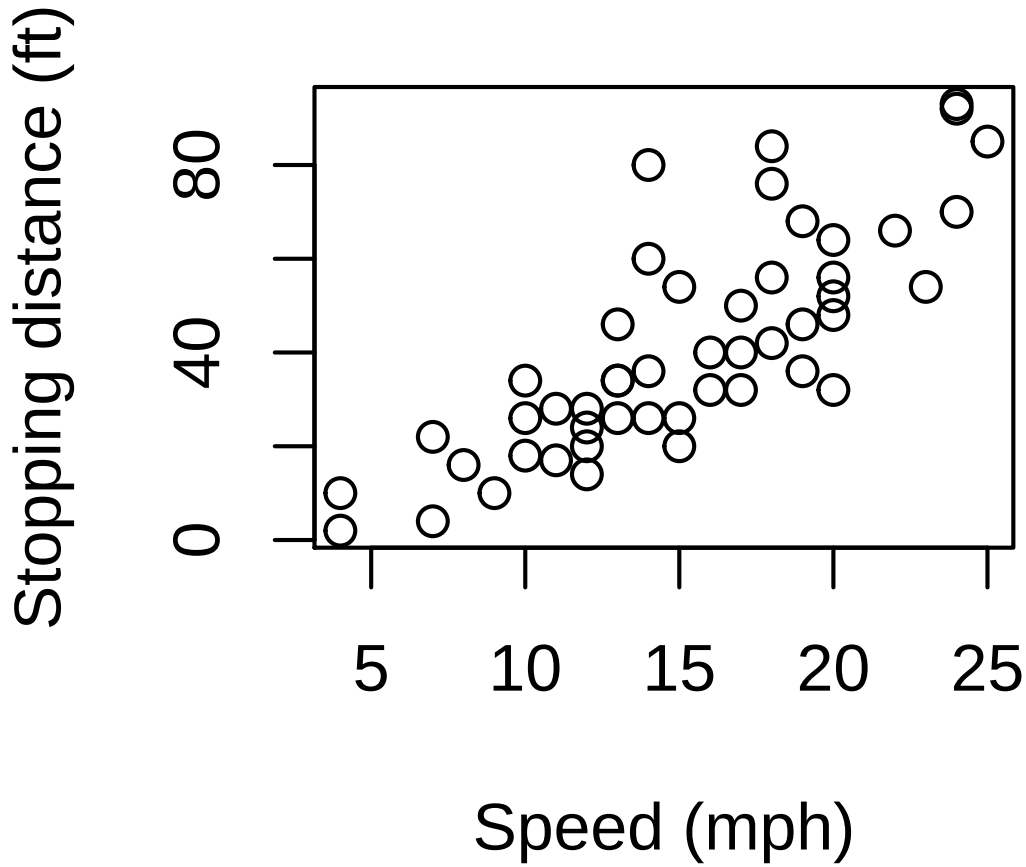
ALL slides for this class have been created using knitr. This means that if R changes, if the dataset changes, if you want to see how I created the slides, *the code and output and interpretation is all there*.

Example 1: how did we make this plot?



Example 1: how did we make this plot?

```
plot(dist ~ speed, data = cars, ylab = "Stopping distance (ft)",
  xlab = "Speed (mph)")
```



Example 2: how did we compute those statistics?

What dataset are they from? If I don't know R, how can I make this table?

```
##      speed           dist
##  Min.   : 4.00   Min.   : 2.00
##  1st Qu.:12.00   1st Qu.:26.00
##  Median :15.00   Median :36.00
##  Mean   :15.22   Mean   :41.41
##  3rd Qu.:19.00   3rd Qu.:56.00
##  Max.   :25.00   Max.   :93.00
```

Example 2: how did we compute those statistics?

What dataset are they from? If I don't know R, how can I make this table?

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.00   Min.   :  2.00  
##  1st Qu.:12.00   1st Qu.: 26.00  
##  Median :15.00   Median : 36.00  
##  Mean   :15.22   Mean   : 41.41  
##  3rd Qu.:19.00   3rd Qu.: 56.00  
##  Max.   :25.00   Max.   : 93.00
```

What is markdown?

- Markdown is a markup language
- Use to write reports, papers, etc.
- Replaces "buttons" in Microsoft Word or similar with code

How to format using markdown

- Preceeding a word with "#" or "##" makes it a header and subheader respectively
- Surrounding a word with "*" or "**" makes it *italics* or **bold** respectively
- Bullets are created by a dash and then a space
- Indented bullets are created by using two spaces before the dash

1. Numbered lists are created by specifying the number, a period, and a space
2. Here is number 2.

How to use knitr

Filling in the content

- Write some text using markdown, e.g. "I am doing an analysis of the iris dataset in R."
- Add code in "code chunks"
  - Start with three backticks and a bracketed {r}
  - End with three backticks

When you are finished, click the knit button to create your dynamic document!

Code chunk options and names

Code chunk names

- Code chunks can be named within the bracket: {r codechunk1} or {r jennascodechunk}
- Useful to indicate what the code chunk does
- Can help with debugging documents
- Can make it easier to navigate a long document

Code chunk options and names

You can specify code chunk options like arguments to an R function

- Code chunk options are specified after a comma following `r` in the code chunk.
- Useful code chunk options:
  - eval: TRUE or FALSE indicating whether you want R to run the code.
  - echo: TRUE or FALSE indicating whether the final document should show the code
  - e.g. {r codechunk1, echo = T, eval = F}

Figures in knitr

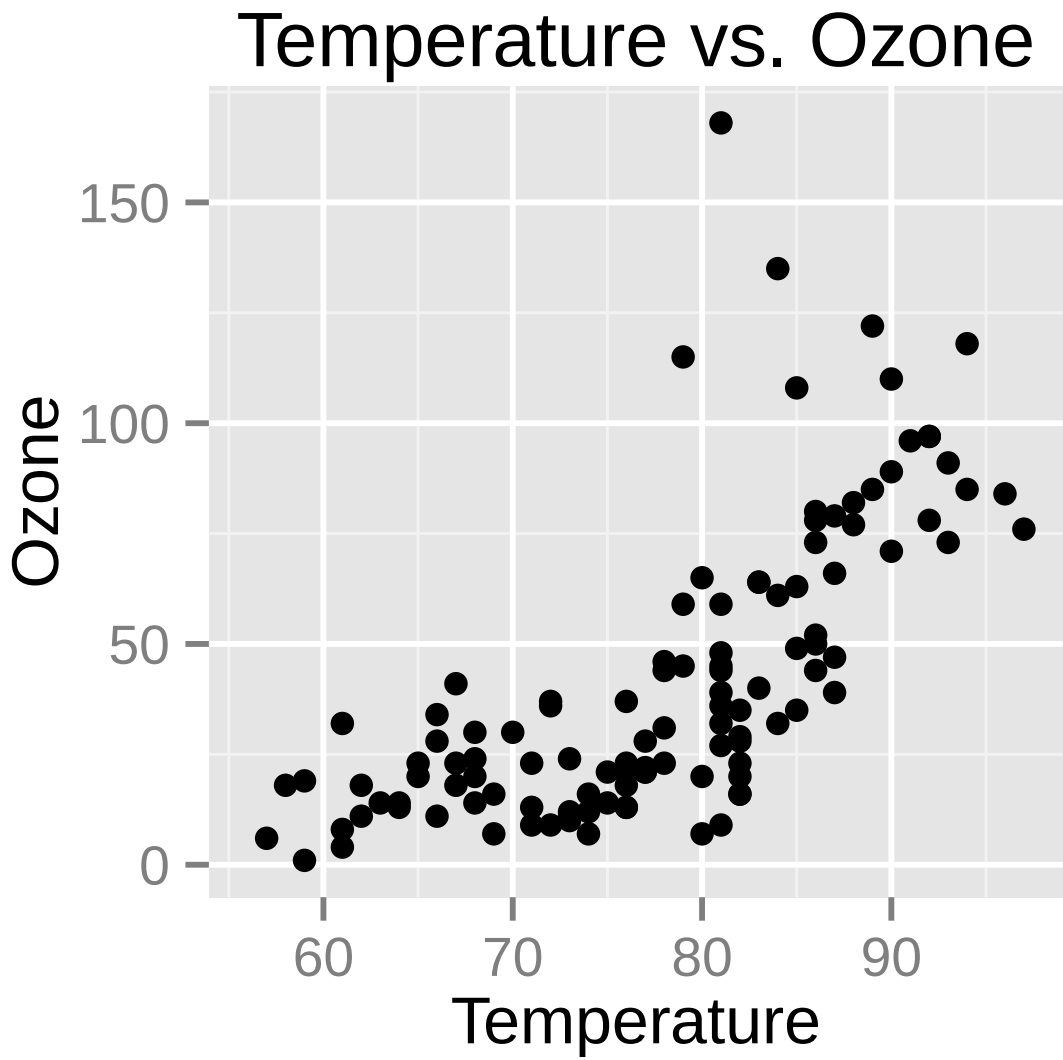
We can include figures in `knitr` using the same commands we have seen earlier.

- You can control the output with code chunk options out.width and out.height, which specify the height and width of the chunk in pixels (for html file)

```
g1 <- ggplot(data = airquality, aes(x = Temp, y = Ozone)) + geom_point() +
    ggtitle("Temperature vs. Ozone") + xlab("Temperature") + ylab("Ozone")
```

Figures in knitr

```
g1
```



Tables in knitr

```
gaq <- group_by(airquality, Month)
s1 <- summarise_each(gaq, funs(mean(., na.rm = T)), Ozone : Temp)
s1
```

```
## Source: local data frame [5 x 5]
##
##   Month    Ozone  Solar.R      Wind     Temp
## 1     5 23.61538 181.2963 11.622581 65.54839
## 2     6 29.44444 190.1667 10.266667 79.10000
## 3     7 59.11538 216.4839  8.941935 83.90323
## 4     8 59.96154 171.8571  8.793548 83.96774
## 5     9 31.44828 167.4333 10.180000 76.90000
```

Tables in knitr

Convenient and easy to use knitr to make tables by using the `kable` function

```
kable(s1)
```

Month	Ozone	Solar.R	Wind	Temp
5	23.61538	181.2963	11.622581	65.54839
6	29.44444	190.1667	10.266667	79.10000
7	59.11538	216.4839	8.941935	83.90323
8	59.96154	171.8571	8.793548	83.96774
9	31.44828	167.4333	10.180000	76.90000

Tables in knitr

We can also use the arguments of kable

```
kable(s1, digits = 2)
```

Month	Ozone	Solar.R	Wind	Temp
5	23.62	181.30	11.62	65.55
6	29.44	190.17	10.27	79.10
7	59.12	216.48	8.94	83.90
8	59.96	171.86	8.79	83.97
9	31.45	167.43	10.18	76.90

Tables in knitr

We can also add captions and realign the output

```
kable(s1, digits = 0, align = rep("c", 4), format = "pandoc",
  caption = "Mean of airquality variables by month")
```

Month	Ozone	Solar.R	Wind	Temp
5	24	181	12	66
6	29	190	10	79
7	59	216	9	84
8	60	172	9	84
9	31	167	10	77

Table 3:Mean of airquality variables by month

In-line code

You can put code in the lines of your text by using backtick-r-space "the code you want" backtick

As an example, here is my code chunk that defines date:

```
date <- Sys.Date()
```

Now I can do inline code specifying the day: Today is 2015-04-01. Now, when I rerun the code on a different day, this will update. Include code that you don't want to run inside two backticks, e.g. today we are learning about `knitr`.

Example 3: Writing interpretations.

```
library(dplyr)
```

The R dataset *cars* has 2 variables which have the labels speed and dist. The dataset has 49 observations. The average speed reported was 15.2244898 miles per hour with a standard deviation (sd) of 5.1932059. The average stopping distance was 41.4081633 (sd = 23.4901738) feet.

Example 3: Writing interpretations.

```
library(dplyr)
cars <- filter(cars, dist < 100)
```

The R dataset *cars* has 2 variables which have the labels speed and dist. The dataset has 49 observations. The average speed reported was 15.2244898 miles per hour with a standard deviation (sd) of 5.1932059. The average stopping distance was 41.4081633 (sd = 23.4901738) feet.

How to use knitr

Creating the document

- Open RStudio and select ‘file-> new file -> R markdown‘
- Give the document a title (title of the document, not the file) and author
- Specify the output (html, pdf, word).
  - We will be using html output for this class

Output files

Rstudio with knitr allows your document to be output as:

1. html
2. pdf
3. word document

Slides using RStudio and knitr

- All slides for this course are made using knitr!
- This means that all code is dynamically included in the document
- This allows me to change the dataset or change a line of code and propagate the results
- It also allows others to look at all the code that I am writing

R presentations using knitr

We can create R presentations using knitr

Creating the presentation

- Open RStudio and select `file-> new file -> R presentation`
- Save the file
- Title followed by string of equal signs indicates a new slide
- Can use markdown and introduce code chunks as in an R Markdown document

Great for teaching!