Literate Statistical Programming with knitr and R Markdown

Biostatistics 140.776

What is knitr?

- An R package written by Yihui Xie (while he was a grad student at Iowa State)
- Available on CRAN
- Supports R Markdown, LaTeX, and HTML as documentation languages
- Can export to PDF, HTML, Word
- Built right into RStudio for your convenience

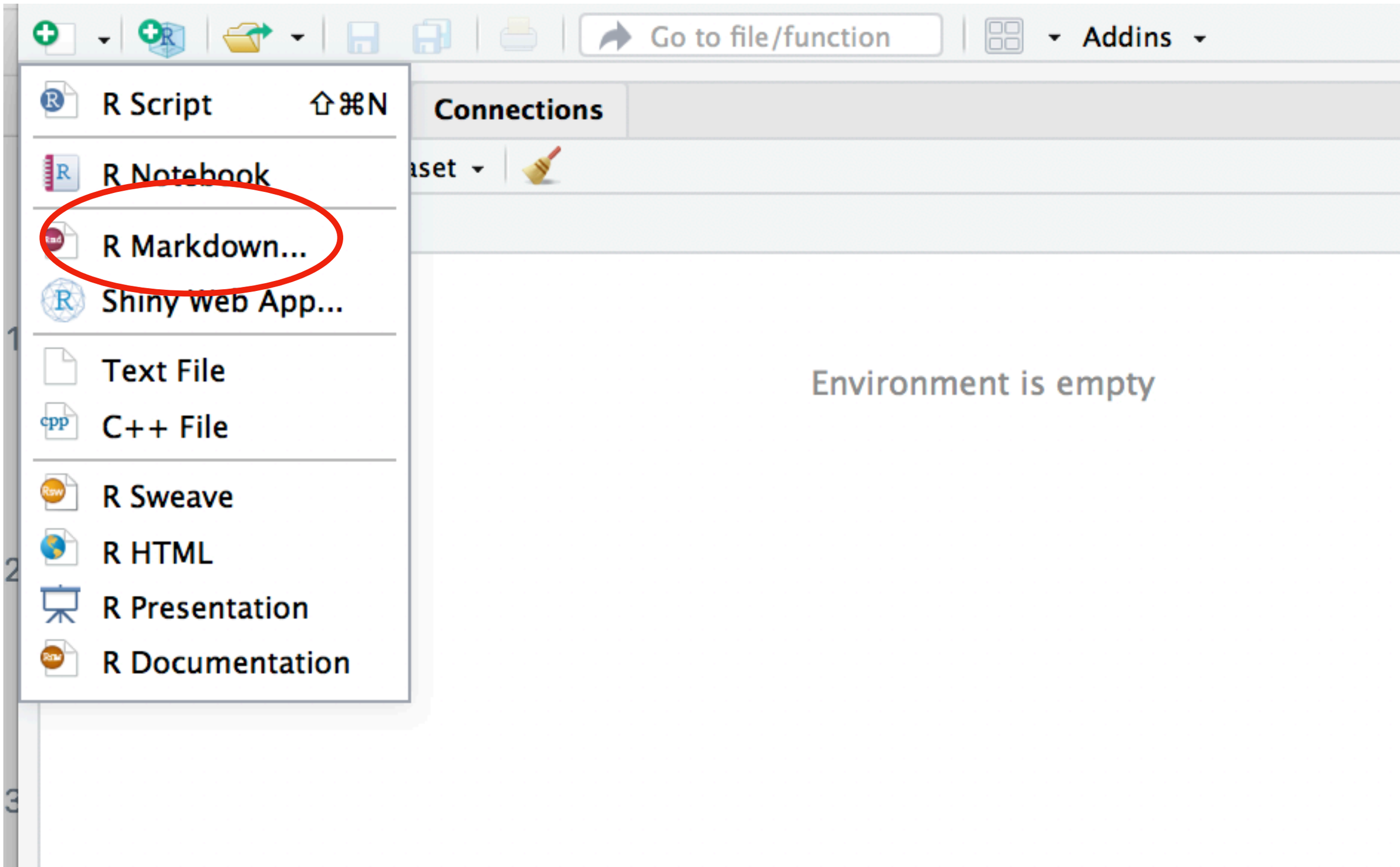
What is knitr Good For?

- Manuals
- Short/medium-length technical documents
- Tutorials
- Reports (esp. if generated periodically)
- Data preprocessing documents/summaries

What is knitr NOT Good For?

- Very long research articles
- Analyses with complex, time-consuming computations
- Documents that require precise formatting
- Documents where formatting needs to be continuously visualized

My First Document!



My First Document!

Untitled1

Knit Insert Run

```
---
title: "Untitled"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```
```

Default Output

Code Chunk

Text Chunk

Code Chunk

Knitting a Document

Untitled1

Knit Insert Run

```
---
title: "Untitled"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```
```

Knitting

```
processing file: Untitled.Rmd
  |.........                                                      |  14%
  ordinary text without R code

  |....................                                            |  29%
label: setup (with options)
List of 1
 $ include: logi FALSE

  |...............................                                 |  43%
  ordinary text without R code

  |.......................................                         |  57%
label: cars
  |...............................................                 |  71%
  ordinary text without R code

  |.......................................................         |  86%
label: pressure (with options)
List of 1
 $ echo: logi FALSE
```

HTML Output

Untitled

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

Code Input

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Code Output

Including Plots

You can also embed plots, for example:

A Few Notes

- knitr will fill a new document with boilerplate text; just delete it
- Code chunks begin with ```{r} and end with ```
- All R code goes in between these markers
- Code chunks can have names, which is useful when we start making graphics

  ```{r firstchunk}
  ## R code goes here
  ```

- By default, code in a code chunk is echoed, as will the results of the computation (if there are results to print)

Processing Documents

- You write the RMarkdown document (.Rmd)
- knitr produces a Markdown document (.md)
- knitr converts the Markdown document into HTML (by default)
- .Rmd —> .md —> .html
- You should NOT edit (or save) the .md or .html documents until you are finished
- By default RStudio does not save the .md document

Another Example

```
# My First knitr Document
Roger D. Peng

## Introduction

This is some text (i.e. a "text chunk").

Here is a code chunk.

```{r, simulation, echo = TRUE}
set.seed(1)
x <- rnorm(100)
mean(x)
```
```

HTML Output

My First knitr Document

Roger D. Peng

Introduction

This is some text (i.e. a “text chunk”).

Here is a code chunk.

```
set.seed(1)
x <- rnorm(100)
mean(x)
```

```
## [1] 0.1088874
```

Hiding the Code

```{r, simulation, echo = FALSE}
set.seed(1)
x <- rnorm(100)
mean(x)
```

HTML Output

My First knitr Document

Roger D. Peng

Introduction

This is some text (i.e. a “text chunk”).

Here is a code chunk.

```
## [1] 0.1088874
```

Inline Computations

```
# My First knitr Document
Roger D. Peng

## Introduction

```{r, computetime, include = FALSE}
time <- format(Sys.time(), "%a %b %d %X %Y")
rand <- rnorm(1)
```

The current time is `r time`.

My favorite random number is `r rand`.
```

Do not show code chunk

Inline Computations

My First knitr Document

Roger D. Peng

Introduction

The current time is Thu Sep 06 08:36:10 2018.

My favorite random number is 1.674726.

Graphics

```
## Introduction

Let's first simulate some data.
```{r, simulatedata, echo = TRUE}
x <- rnorm(100)
y <- x + rnorm(100, sd = 0.5)
```

Here is a scatterplot of the data.
```{r, scatterplot, fig.height = 4}
library(ggplot2)
qplot(x, y, main = "My Simulated Data")
```
```

Adjust figure height

Graphics

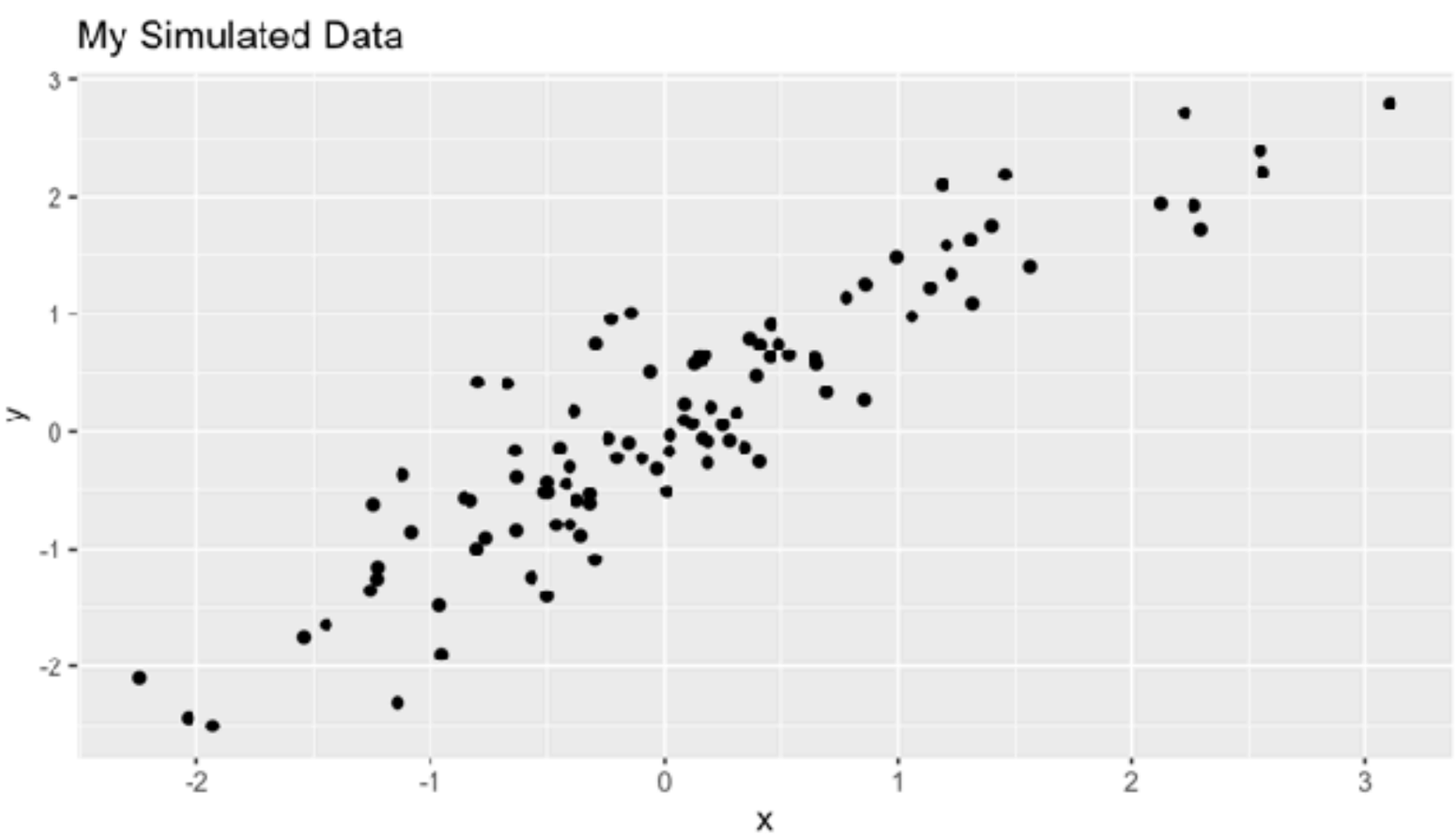
Introduction

Let's first simulate some data.

```
x <- rnorm(100)
y <- x + rnorm(100, sd = 0.5)
```

Here is a scatterplot of the data.

```
library(ggplot2)
qplot(x, y, main = "My Simulated Data")
```



Graphic (embedded)

Tables

```
## Data Summary
```

```{r, loaddata, include = FALSE}
library(datasets)
data(airquality)
```

```{r,summary}
library(tableone)
tab <- CreateTableOne(c("Ozone", "Wind", "Temp", "Solar.R"),
                      data = airquality)
summary(tab)
```

Tables

Data Summary

```
library(tableone)
tab <- CreateTableOne(c("Ozone", "Wind", "Temp", "Solar.R"),
                      data = airquality)
summary(tab)
```

```
##
##      ### Summary of continuous variables ###
##
## strata: Overall
##           n miss p.miss mean sd median p25 p75 min max skew kurt
## Ozone   153   37     24   42 33     32  18  63   1 168  1.2  1.3
## Wind    153    0      0   10  4     10   7  12   2  21  0.3  0.1
## Temp    153    0      0   78  9     79  72  85  56  97 -0.4 -0.4
## Solar.R 153    7      5  186 90    205 116 259   7 334 -0.4 -1.0
```

Tables (Formatted)

```
## Data Summary
```

```{r, loaddata, include = FALSE}
library(datasets)
data(airquality)
```

```{r,summary,results="asis"}
library(xtable)
print(xtable(airquality), type = "html")
```

Tables (Formatted)

Data Summary

```
library(xtable)
print(xtable(airquality), type = "html")
```

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.40	67	5	1
2	36	118	8.00	72	5	2
3	12	149	12.60	74	5	3
4	18	313	11.50	62	5	4
5			14.30	56	5	5
6	28		14.90	66	5	6
7	23	299	8.60	65	5	7
8	19	99	13.80	59	5	8
9	8	19	20.10	61	5	9
10		194	8.60	69	5	10
11	7		6.90	74	5	11
12	16	256	9.70	69	5	12

Setting Global Options

- Sometimes we want to set options for every code chunk that are different from the defaults
- For example, we may want to suppress all code echoing and results output
- We have to write some code to set these global options (usually at the beginning of the document)

Global Options

```{r, include = FALSE}
knitr::opts_chunk$set(echo = FALSE)
```

Introduction

First simulate some data.

```{r, simulatedata}
x <- rnorm(100)
y <- x + rnorm(100, sd = 0.5)
```

Here's a scatterplot.

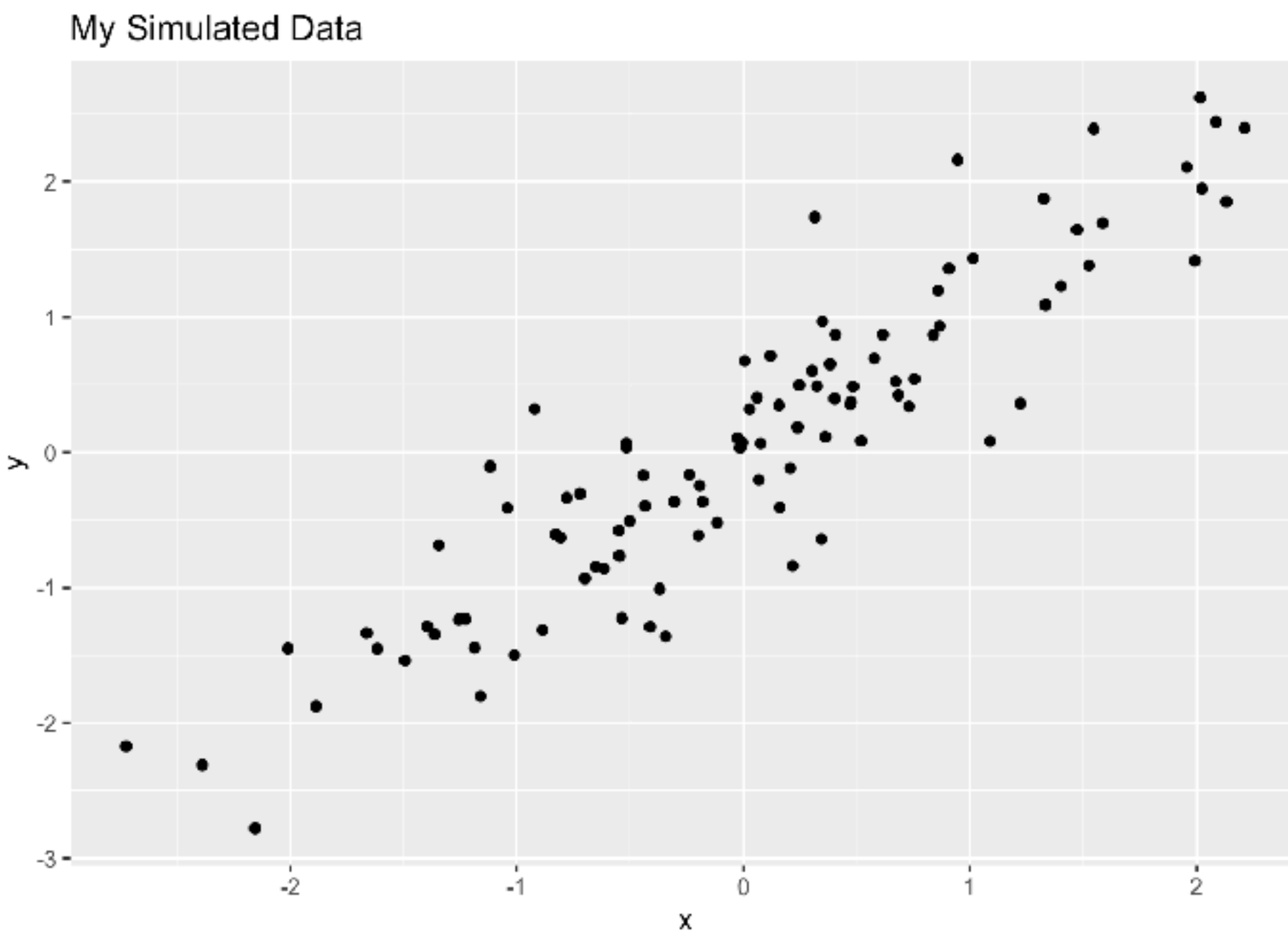
```{r, scatterplot}
library(ggplot2)
qplot(x, y, main = "My Simulated Data")
```

Global Options

Introduction

First simulate some data.

Here's a scatterplot.



Overriding Global Options

```{r, include = FALSE}
knitr::opts_chunk$set(echo = FALSE)
```

Introduction

First simulate some data.

```{r, simulatedata, echo = TRUE}
x <- rnorm(100)
y <- x + rnorm(100, sd = 0.5)
```

Here's a scatterplot.

```{r, scatterplot}
library(ggplot2)
qplot(x, y, main = "My Simulated Data")
```

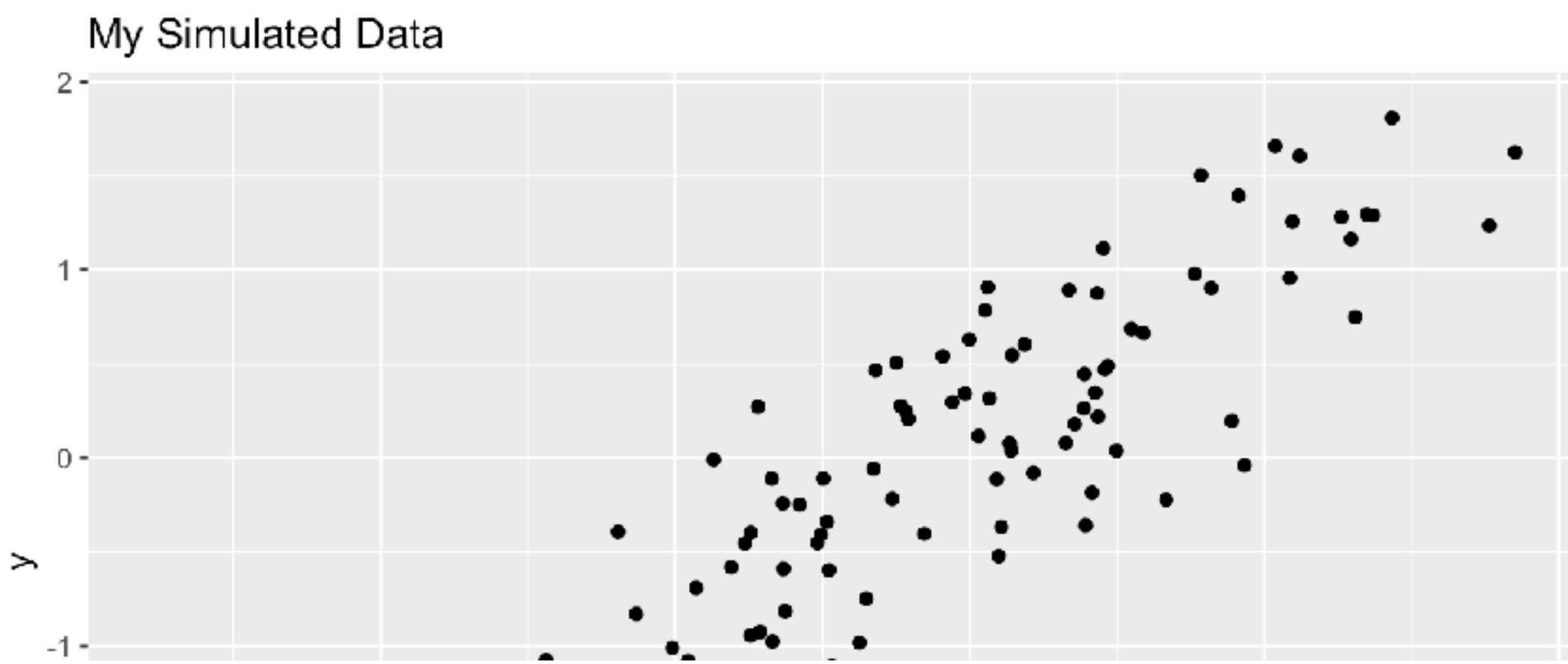
Overriding Global Options

Introduction

First simulate some data.

```
x <- rnorm(100)
y <- x + rnorm(100, sd = 0.5)
```

Here's a scatterplot.



Caching Computations

- What if one chunk takes a long time to run?
- All chunks have to be re-computed every time you re-knit the file
- The `cache=TRUE` option can be set on a chunk-by-chunk basis to store results of computation
- After the first run, results are loaded from cache

Caching Computations

- If the data or code (or anything external) changes, you need to re-run the cached code chunks
- Dependencies are not checked explicitly
- Chunks with significant side effects may not be cacheable

Summary

- Literate statistical programming can be a useful way to put text, code, data, output all in one document
- knitr is a powerful tool for integrating code and text in a simple document format
- Particularly useful for “work-in-progress” reports and for regularly generated monitoring-style output
- Code + output can be overwhelming so be judicious with what you show others!