Teaching Introductory Statistics with R

Discussion

Alison L. Gibbs
alison.gibbs@utoronto.ca

Department of Statistical Sciences
University of Toronto

SSC 2014

Choices...

Menu-driven

- Minitab
- StatCrunch
- SPSS
- ...

Command-driven

- SAS
- Matlab
- R
- ...

Choices...

Menu-driven

- Minitab
- StatCrunch
- SPSS
- ...

Command-driven

- SAS
- Matlab
- R
- ...

1. Which do you use?

Choices...

Menu-driven

- Minitab
- StatCrunch
- SPSS
- ...

Command-driven

- SAS
- Matlab
- R
- ...

1. Which do you use?
2. Which do you expect that users from other disciplines or industry will use?

Choices...

Menu-driven

- Minitab
- StatCrunch
- SPSS
- ...

Command-driven

- SAS
- Matlab
- R
- ...

1. Which do you use?
2. Which do you expect that users from other disciplines or industry will use?
3. Which do you think are useful for your students?

R for statistics novices

It can be done!

R for statistics novices

It can be done!

- *Joel's class:* distance-education, professional Masters
- *Kevin's class:* MPH epidemiology program (with undergraduate course in statistics)

R for statistics novices

It can be done!

- *Joel's class:* distance-education, professional Masters
- *Kevin's class:* MPH epidemiology program (with undergraduate course in statistics)

Could it also work for junior undergraduates?

R for statistics novices

It can be done!

- *Joel's class:* distance-education, professional Masters
- *Kevin's class:* MPH epidemiology program (with undergraduate course in statistics)

Could it also work for junior undergraduates?

An email from a colleague in biology

Hi Alison,
I wanted to enquire whether you might know if Stats is considering changing to R in the intro course... I have found that our undergrad research students with some experience with R seem much more capable to jump into analysis of their own project data, and I suspect that other disciplines also frequently use R. I don't know what your position might be on this issue, but I know from the perspective of biology that it would be really helpful if students learned R as a practical element of their intro stats training.

"Are we crazy?"

hist {graphics} R Documentation

Histograms

Description

The generic function `hist` computes a histogram of the given data values. If `plot = TRUE`, the resulting object of class `"histogram"` is plotted by `plot.histogram`, before it is returned.

Usage

```
hist(x, ...)

## Default S3 method:
hist(x, breaks = "Sturges",
     freq = NULL, probability = !freq,
     include.lowest = TRUE, right = TRUE,
     density = NULL, angle = 45, col = NULL, border = NULL,
     main = paste("Histogram of" , xname),
     xlim = range(breaks), ylim = NULL,
     xlab = xname, ylab,
     axes = TRUE, plot = TRUE, labels = FALSE,
     nclass = NULL, warn.unused = TRUE, ...)
```

Arguments

`x`	a vector of values for which the histogram is desired.

Could R be the right choice for my introductory course?

- Yes! If you want it to be.

Could R be the right choice for my introductory course?

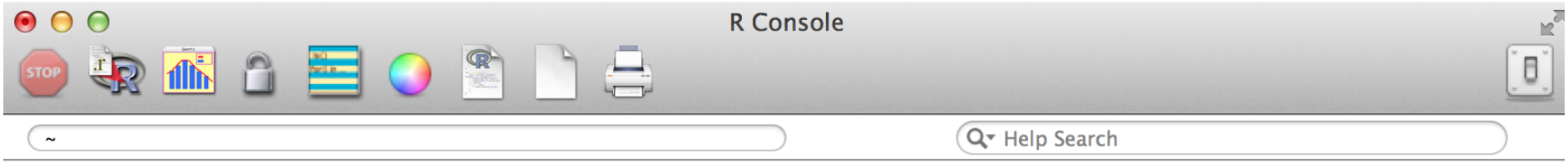
- Yes! If you want it to be.
- Your students won't outgrow it.

Could R be the right choice for my introductory course?

- Yes! If you want it to be.
- Your students won't outgrow it.
- There are tools to make it easier for novices.

Adding training wheels

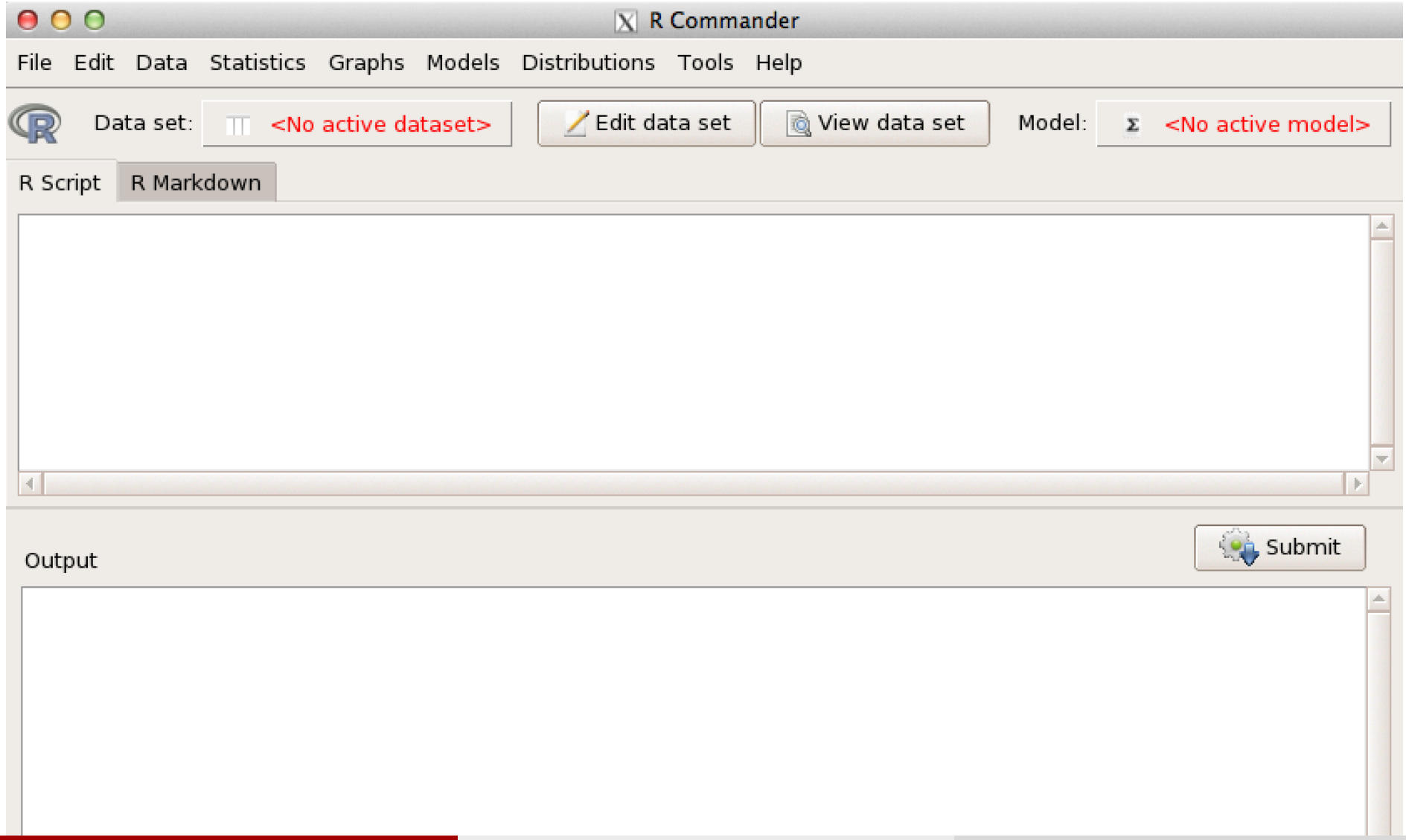
Instead of ....




```
R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin10.8.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.62 (6558) x86_64-apple-darwin10.8.0]

[History restored from /Users/alisong/.Rapp.history]

>
```

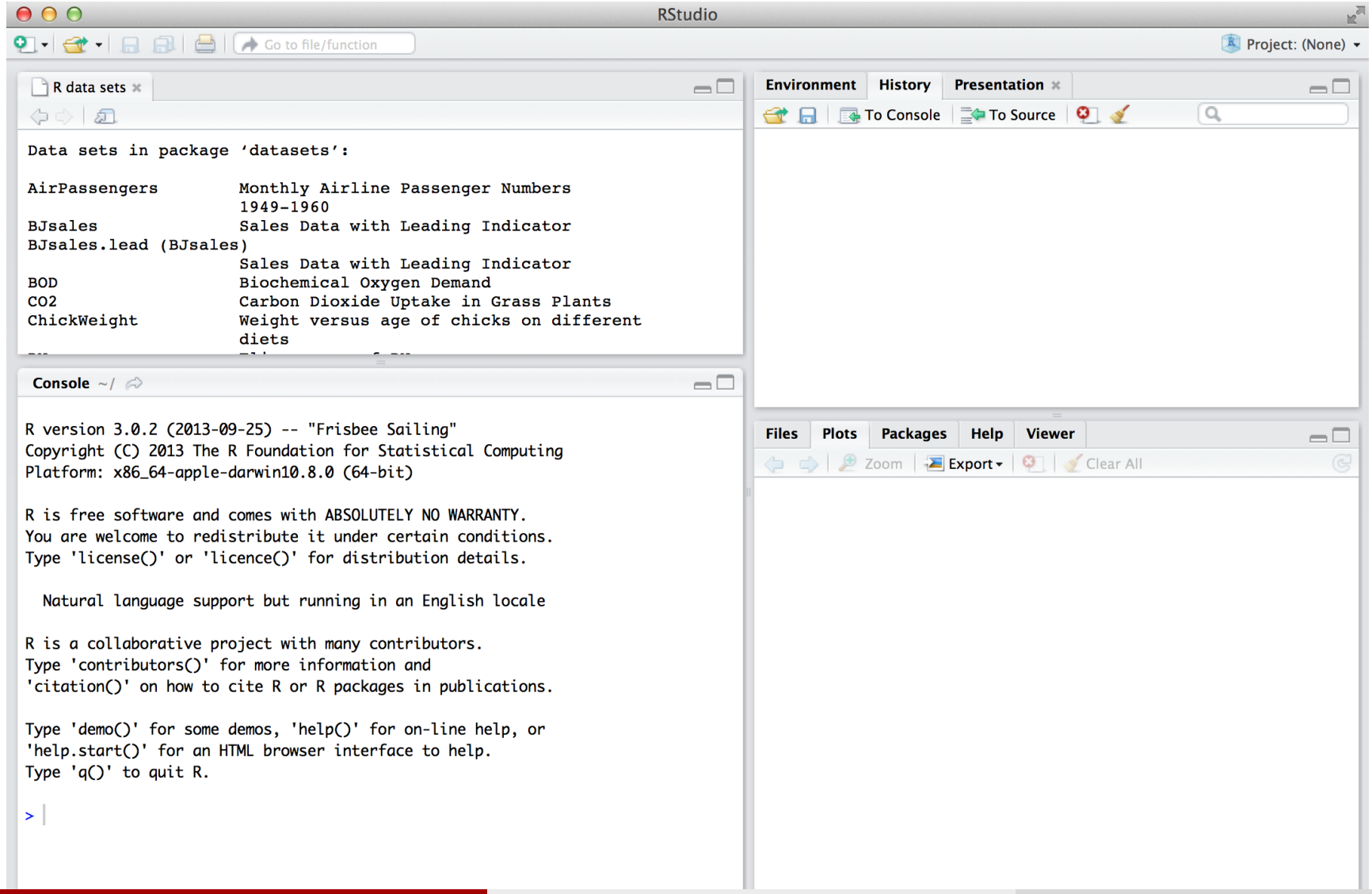
Adding training wheels

How about ....?



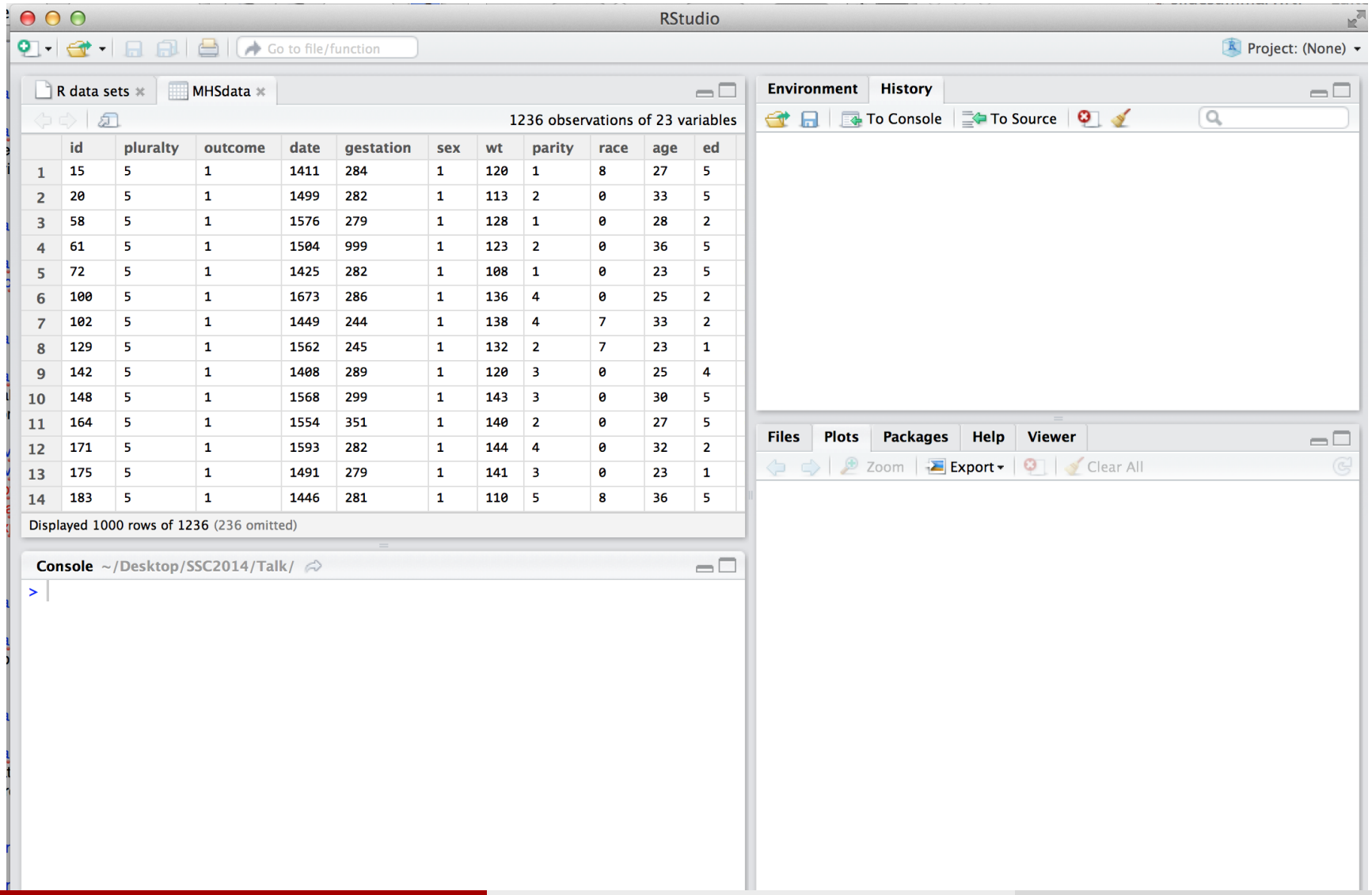
Adding training wheels

Or ....?



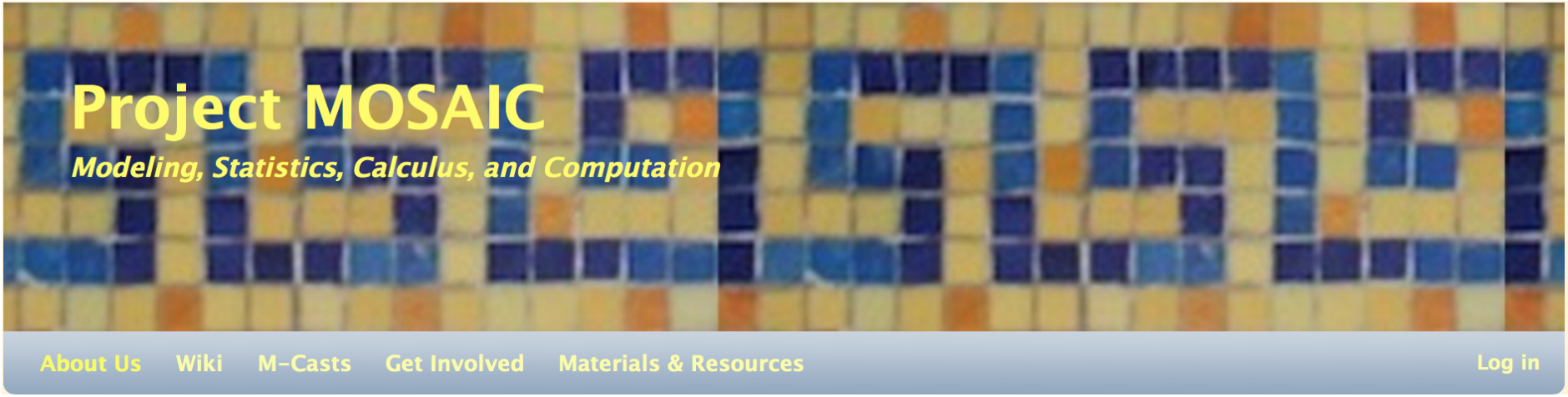
Adding training wheels

Or RStudio on a server with the data pre-loaded?



	id	pluralty	outcome	date	gestation	sex	wt	parity	race	age	ed
1	15	5	1	1411	284	1	120	1	8	27	5
2	20	5	1	1499	282	1	113	2	0	33	5
3	58	5	1	1576	279	1	128	1	0	28	2
4	61	5	1	1504	999	1	123	2	0	36	5
5	72	5	1	1425	282	1	108	1	0	23	5
6	100	5	1	1673	286	1	136	4	0	25	2
7	102	5	1	1449	244	1	138	4	7	33	2
8	129	5	1	1562	245	1	132	2	7	23	1
9	142	5	1	1408	289	1	120	3	0	25	4
10	148	5	1	1568	299	1	143	3	0	30	5
11	164	5	1	1554	351	1	140	2	0	27	5
12	171	5	1	1593	282	1	144	4	0	32	2
13	175	5	1	1491	279	1	141	3	0	23	1
14	183	5	1	1446	281	1	110	5	8	36	5

Training wheels plus a downhill push: *The mosaic package*



- make it easier to learn R by minimizing and simplifying the coding
- with consistent syntax for function calls

$$\boxed{\text{what I want}}\left(\boxed{\text{Y}} \sim \boxed{\text{X}},\ \textit{data} = \boxed{\text{my data}}\right)$$

- only the essential functions
- protect students from vectors and data frames (for now)

Promoting Reproducible Research

Promoting Reproducible Research

The Economist

Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

Promoting Reproducible Research

The Economist

Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

---

The New York Times | http://nyti.ms/1f6l0Kq

SCIENCE

New Truths That Only One Can See

Promoting Reproducible Research

nature International weekly journal of science

ANNOUNCEMENT

Reducing our irreproducibility

Over the past year, *Nature* has published a string of articles that highlight failures in the reliability and reproducibility of published research (collected and freely available at go.nature.com/huhbyr). The problems arise in laboratories, but journals such as this one compound them when they fail to exert sufficient scrutiny over the results that they publish, and when they do not publish enough information for other researchers to assess results properly.

From next month, *Nature* and the Nature research journals will introduce editorial measures to address the problem by improving the consistency and quality of reporting in life-sciences articles. To ease the interpretation and improve the reliability of published results we will more systematically ensure that key methodological details are reported, and we will give more space to methods sections. We will examine statistics more closely and encourage authors to be transparent, for example by including their raw data.

Central to this initiative is a checklist intended to prompt authors to disclose technical and statistical information in their submissions, and to encourage referees to consider aspects important for research reproducibility (go.nature.com/oloeip). It was developed after discussions with researchers on the problems that lead to irreproducibility, including workshops organized last year by US National Institutes of Health (NIH) institutes. It also draws on pub-

we will commission statisticians as consultants on certain papers, at the editor's discretion and at the referees' suggestion.

We recognize that there is no single way to conduct an experimental study. Exploratory investigations cannot be done with the same level of statistical rigour as hypothesis-testing studies. Few academic laboratories have the means to perform the level of validation required, for example, to translate a finding from the laboratory to the clinic. However, that should not stand in the way of a full report of how a study was designed, conducted and analysed that will allow reviewers and readers to adequately interpret and build on the results.

To allow authors to describe their experimental design and methods in as much detail as necessary, the participating journals, including *Nature*, will abolish space restrictions on the methods section.

To further increase transparency, we will encourage authors to provide tables of the data behind graphs and figures. This builds on our established data-deposition policy for specific experiments and large data sets. The source data will be made available directly from the figure legend, for easy access. We continue to encourage authors to share detailed methods and reagent descriptions by depositing protocols in Protocol Exchange (www.nature.com/protocolexchange), an open resource linked from the primary paper.

Renewed attention to reporting and transparency is a small step. Much bigger underlying issues contribute to the problem, and are beyond the reach of journals alone. Too few biologists receive adequate training in statistics and other quantitative aspects of their subject. Mentoring of young scientists on matters of rigour and transparency is inconsistent at best. In academia, the ever increas-

Promoting Reproducible Research with R Markdown

- Data analysis is an iterative process.
- Can we capture that in a way that is comprehensible to someone else?
- And reproducible?
  - Easy to reproduce, update, share
- Suggestion: Use R Markdown to integrate code, output, commentary

Promoting Reproducible Research with R Markdown

- Data analysis is an iterative process.
- Can we capture that in a way that is comprehensible to someone else?
- And reproducible?
  - Easy to reproduce, update, share
- Suggestion: Use R Markdown to integrate code, output, commentary

Text block
R chunk
Text block
R chunk

Knit together as text, code, and output, including plots.

Promoting Reproducible Research with R Markdown

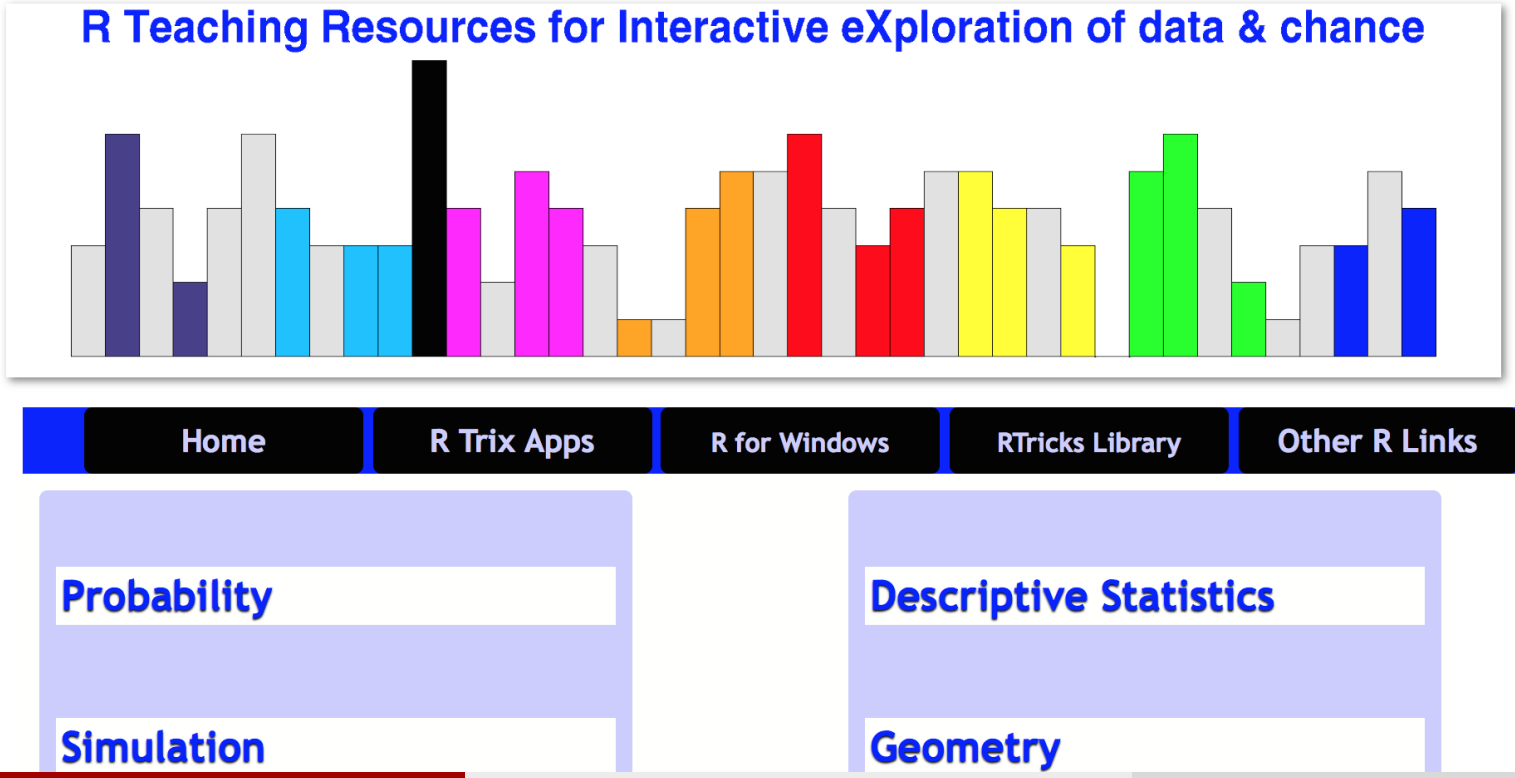
- One of Kevin's assignments: write up the analysis "like a standard scientific paper."
- Joel's challenge: how to assess the use of R.
- Kevin noted the danger of having learning the package obscure learning the statistics.

Promoting Reproducible Research with R Markdown

- One of Kevin's assignments: write up the analysis "like a standard scientific paper."
  Train them early with reproducibility in mind.
- Joel's challenge: how to assess the use of R.
  Hand in a document integrating executable code, output, and commentary.
- Kevin noted the danger of having learning the package obscure learning the statistics.
  But maybe the package can help us capture the statistical process.

R for all ages

http://www.stats.uwo.ca/faculty/braun/RTricks/RTrixApps.php



The final word ...

"Not only can R be integrated into introductory (biostatistics) courses, but it can enhance statistical understanding."