Day 1, Part 1: R Introduction

Brennan Terhune-Cotter and Matt Dye

https://github.com/brennangitsit/2023_IAM3_R | Thanks to Adam Stone for creating and sharing the content in this slidedeck!

Welcome to IAM3!

Intensive
Advanced
PrograMming
3x harder than normal programming

Welcome to R!

R is a user-friendly, intuitive programming language that allows researchers to efficiently manage, analyze, and visualize data.

R is widely used and has become the de facto language for data management and analysis in the scientific community; thus, it is important to learn and can be immensely helpful for you in the future!

Why Use R?

R is difficult and time-consuming to learn. However, learning R will be worth your while if you want to become a pro at data analysis!

The benefits of processing your data via R include:

- You can wrangle and analyze your data more efficiently and without mistakes from manual data entry.
- You can re-run your analysis as many times as you want!
- You can easily share your analysis with others by sending them your scripts.
- Your analysis is fully *reproducible* - you have an exact record of what you did with your data from start to finish. ¶

Reproducibility

- Reproducible science is the name of the game![1]

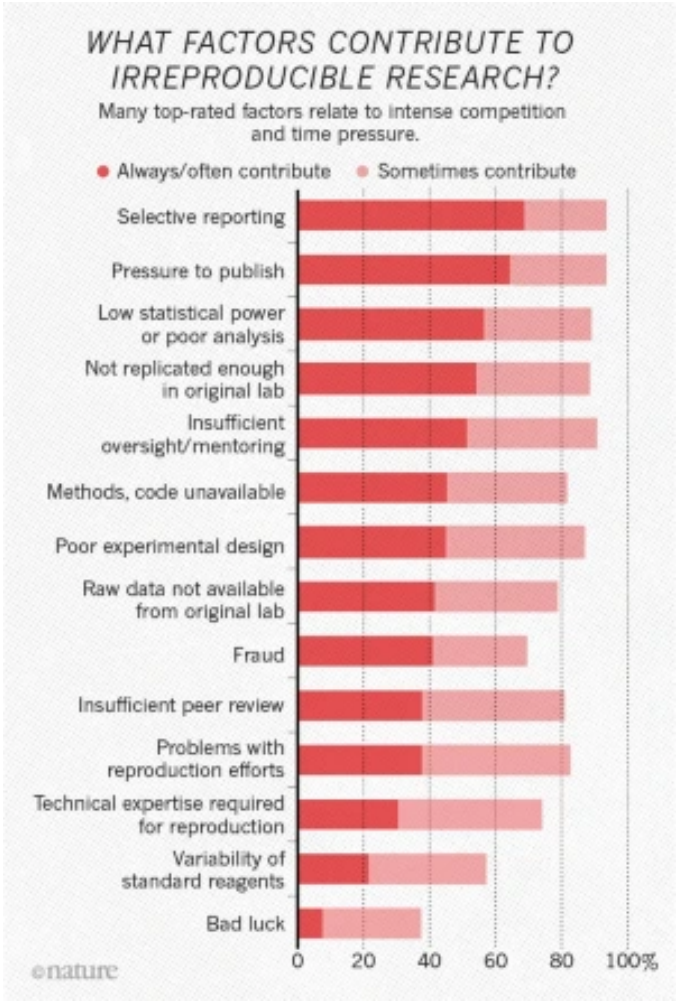
- **Methods reproducibility:** Same procedures on data can be exactly repeated by the same team
- **Results reproducibility:** Same results can be obtained from an independent study with the same procedures
- **Inferential reproducibility:** Same conclusions can be drawn from an independent replication or a reanalysis
- Analysis via R addresses **methods reproducibility** & can help with other forms of reproducibility. ¶

1. Plesser, H. E. (2017). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. Frontiers in Neuroinformatics, 11, 76.

Reproducibility

Using reproducible workflows via R will mitigate/solve:

1. Poor statistical analysis
2. Unavailable methods/code
3. Unavailable raw data



Advantages of Coding in R

1. Code is independent of data
2. Code handles large or in-progress datasets
3. Code is documentation
4. Code outlives your research team
5. Code generates the visualizations that you actually want
6. R has an amazing community & resources
7. You become a coder!

1. Code is independent of data

- **Your data is sacrosanct.** You shouldn't ever mess with it.
- Code lets you analyze data without touching it.
- It follows that we shouldn't be analyzing data using spreadsheets. ¶

What Not to Do

JPMorgan Discloses $2 Billion in Trading Losses

BY JESSICA SILVER-GREENBERG AND PETER EAVIS MAY 10, 2012 10:11 PM 421

Mario Tama/Getty Images Jamie Dimon, the chief executive of JPMorgan Chase.

JPMorgan Chase, which emerged from the financial crisis as the nation's biggest bank, disclosed on Thursday that it had lost more than $2 billion in trading, a surprising stumble that promises to

What Not to Do

Caution

““operated through a series of Excel spreadsheets, which had to be completed manually, by a process of copying and pasting data from one spreadsheet to another”, and “that it should be automated” but never was.”

Caution

“After subtracting the old rate from the new rate, the spreadsheet divided by their sum instead of their average, as the modeler had intended.” [1]

1. https://blog.revolutionanalytics.com/2013/02/did-an-excel-error-bring-down-the-london-whale.html

What Not to Do, Part Two

Growth in a **Time** of **Debt**

CM Reinhart, KS Rogoff - American economic review, 2010 - pubs.aeaweb.org

… relationship between high public **debt** levels, **growth** and inflation.Our … **growth** and **debt** seems relatively weak at "normal" **debt** levels, median **growth** rates for countries with public **debt** …

☆ Save Cite Cited by 4887 Related articles All 55 versions

1,730 | 57 | 1,288 | 14

JOHN CASSIDY

THE REINHART AND ROGOFF CONTROVERSY: A SUMMING UP

By John Cassidy
April 26, 2013

In one of life's little ironies, last Friday's disappointing G.D.P. figures, which reflected a sharp fall in government spending, appeared on the same day that the economists Carmen Reinhart and Kenneth Rogoff published an Op-Ed in the *Times* defending their famous (now infamous) research that conservative politicians around the world had seized upon to justify penny-pinching policies. Addressing a new paper by three lesser lights of their profession from the University of Massachusetts, Amherst, which uncovered data omissions, questionable methods of weighting, and elementary coding errors in Reinhart and Rogoff's original work, and which went around the

	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5	n.a.	7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

Cleaning Up Data Manually

Suppose you have survey data from 16 deaf subjects. It looks like this…

subjid	age_deaf	age_exposed_asl	learn_asl_from
1	0	0	Parents, Siblings, Friends, Teachers
2	0	0.2	Parents, Friends, Teachers
3	0	birth	Parents, Siblings, Friends, Teachers, Other
4	0	1	Friends, Teachers
5	0	birth	Friends, Other
6	0	2 years old or older	Siblings, Friends, Teachers

Cleaning Up Data Manually

This data needs to be cleaned up before analysis. You could do this manually via Excel. To be able to get a mean ASL age of exposure, you:

- replace “birth” with “0”
- replace “2 years old or older” with “2”

However:

- you have no *documentation* of what you did (unless you write it down manually)
- other people *cannot see* what decisions you made
- if you get more data, you will have to manually recode the data again.
- This leads to mistakes, especially if the project spans multiple researchers and/or years! ¶

Cleaning Up Data via R

Alternatively, you could use R:

```
library(tidyverse) # for pipe (i.e., %>% )
library(dplyr) # for mutate() and case_when()
cleaner_data <- messy_data %>%
  mutate(age_exposed_asl = case_when(age_exposed_asl == "birth" ~ "0",
                                     age_exposed_asl == "2 years old or older" ~ "2",
                                     TRUE ~ age_exposed_asl
                                     )
         )
```

This code uses the *mutate()* and *case_when()* functions in the *dplyr* package to replace the two non-numeric values while keeping the rest of the values the same.

Cleaning Up Data via R

See the results!

subjid	age_deaf	age_exposed_asl	learn_asl_from
1	0	0	Parents, Siblings, Friends, Teachers
2	0	0.2	Parents, Friends, Teachers
3	0	0	Parents, Siblings, Friends, Teachers, Other
4	0	1	Friends, Teachers
5	0	0	Friends, Other
6	0	2	Siblings, Friends, Teachers

2. Code handles in-progress datasets

- You can write the code before you collect all of your data
- Then you add new files to your project folder, and rerun your code
- The code will find the new data and update everything! ¶

Cleaning up Data via R

Suppose a few months later you add two more subjects:

subjid	age_deaf	age_exposed_asl	learn_asl_from
7	0	0	Parents, Siblings, Friends, Teachers
8	0	birth	Parents, Siblings, Extended family (e.g. grandparents, aunts, uncles), Friends, Teachers

Instead of manually recoding the data, and hoping it stays consistent from the last time, you can simply rerun the script on the whole dataset to get:

subjid	age_deaf	age_exposed_asl	learn_asl_from
7	0	0	Parents, Siblings, Friends, Teachers
8	0	0	Parents, Siblings, Extended family (e.g. grandparents, aunts, uncles), Friends, Teachers

R allows you to code your data *once* and apply it consistently to *any dataset you have*, including future data!

3. Code is documentation

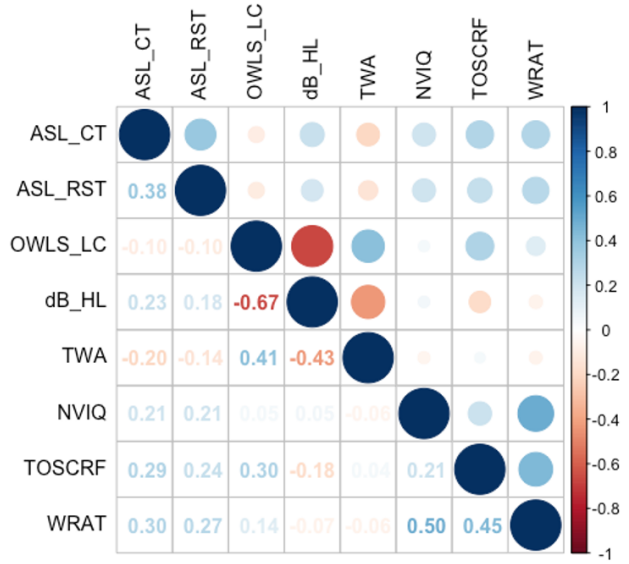
- R code is *self-documenting*!
- Well-written and commented code tells you what you did with the data.
- You can always go back and see what the code does if you're not sure what you did. ¶

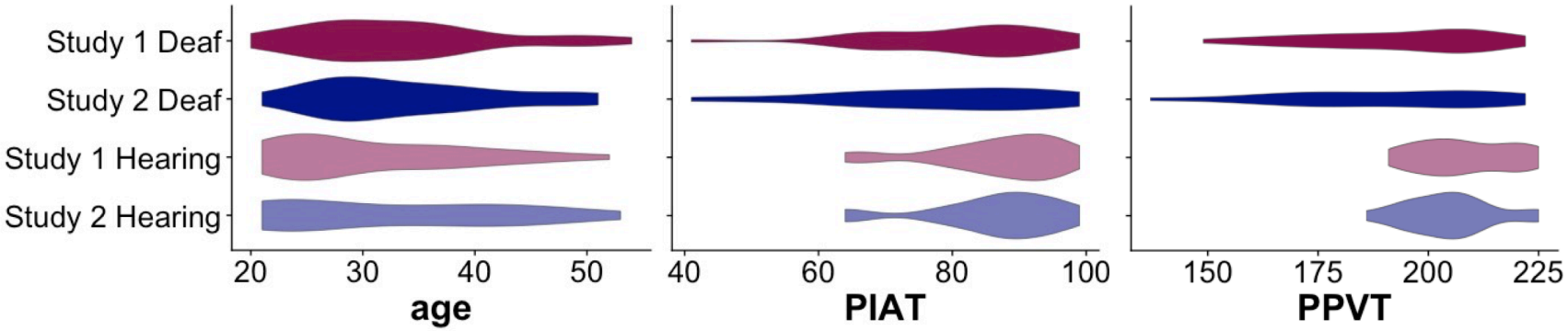
4. Code outlives your research team

- Code allows consistency across *researchers* as well as *subjects* and *time*.
- Can you be sure your RAs are doing manual work right every single time? Are they doing it the same way as previous RAs?
- Code automates the manual and repetitive work that humans often make mistakes on. ¶

5. Code generates the visualizations you want

- You can use ggplot2 in R to generate a stunning array of visualizations, limited only by your imagination! 🤩
- No more wrestling with Excel or Google charts 🖐



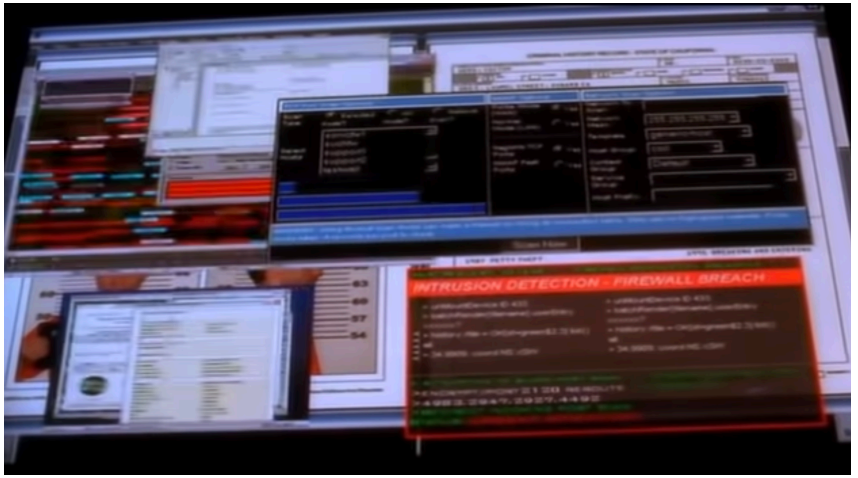


6. R has an amazing community & resources

- If you have a problem, other people have had & solved that problem before.
- There are open-source packages for almost everything under the sun.
- There is extensive documentation for everything in R.
- StackExchange and ChatGPT work really well for solving problems with R.
- Hadley Wickham has a free and well-written R textbook.
- These resources will be covered in the last day. ¶

7. You become a coder!

- Welcome to the world of computer programming!
- R is a great way to get started via a user-friendly language which can be applied to things you actually use (i.e., data). ¶



What to expect...

Structure of this workshop

- This workshop will teach you R in an accessible way (coming from someone who only took one formal programming class and none in R!).
- The primary purpose of this workshop is to teach you the basics of R programming *and get you started on using R for your own research*.
- Some of you have more experience with R than others, so:
  - We will be giving lectures half of the time.
  - The other half will be lab time where you can:
    - practice the concepts we've taught *or*
    - work on improving your own R skills using your own data, with our help (OYOLabs).
  - Sometimes we'll split sessions into *review labs* and *"advanced" lectures*. ¶

On Your Own (OYO) Labs

- The hardest part of learning R is applying it to your own specific situation and your own data/projects.
- OYOLabs are meant for you to build your own workspace (or workspaces) with R so that you have a foundation to continue using R with your own projects after the Institute
  - You will have Brennan, Matt, and each other to help you work through the problems that will inevitably crop up!
- We encourage you to use the OYOLabs to work on your projects *as they relate to that day*. For example, on Tuesday (data visualization day!) you should try to write or refine scripts which visualize your data.
  - However, if you wish to work on other aspects of your R project instead, use your time the best way you feel fit! ¶

R Clinic

- We will have an *R Clinic* on Saturday afternoon.
- This is time for you to meet with me and/or Matt for help and advice with your analysis.
- Length of appointments will depend on how many people are interested.
- You are encouraged to also come up to us any other time during the Institute (such as lunchtime) if you desire answers, help, conversation, or a sympathetic shoulder for gentle weeping! ¶

Workshop schedule (in COURSE OVERVIEW)

Time	AM	PM	Monday **June 12th**	Tuesday **June 13th**	Wednesday **June 14th**	Thursday **June 15th**	Friday **June 16th**	Saturday **June 17th**
15m		**14:00**		Review				R Clinic
45m		**14:15 | 15:00**		Slides 1: R Basics				
10m	**10:00 | 10:10**	**15:00 | 15:10**		Break	Review	Review	Review	
30m	**10:10 | 10:40**	**15:10 | 15:40**	Slides 1: Why Use R?	Lab 1: R Basics	Slides 1: Import and Transform Data using Tidyverse	Slides 1: Tidy and Wrangle Data	Slides 1: Summarize and Report Data	
5m	**10:40**	**15:40**	Break	Break	Break	Break	Break	
20m	**10:45 | 11:05**	**15:45 | 16:05**	Slides 2: R Env & RStudio	Slides 2: Data Viz / *Slides 3: Adv. Viz*	Lab 1	Lab 1	Lab 1	
5m	**11:05**	**16:05**	Break	Break	Break	Break	Break	
20m	**11:10 | 11:30**	**16:10 | 16:30**	Slides 3 and Lab: Import Your Data	Lab 2: Viz / *Lab 3: Adv. Viz*	Review Lab / *Slides 2: Trans. Data*	Review Lab / *Slides 2: Manage Your Project*	Slides 2: Resources for Learning R	
30m	**11:30 | 12:00**	**16:30 | 17:00**	OYOLab: Set up your project	OYOLab	OYOLab	OYOLab	OYOLab	