

# Statistical analysis

[R]

Dani Arribas-Bel & Thomas De Graaff

September 5, 2014

# Outline

# Today

- ▶ Reproducible statistical analysis
- ▶ Reinhart & Rogoff: a textbook example of the power of replication
- ▶ R: what is it and why should I care?
- ▶ R overview
  - ▶ Libraries and help
  - ▶ Reading data
  - ▶ Exploring the `data.frame`
  - ▶ Manipulate a `data.frame`
  - ▶ Analyze data
  - ▶ Visualize data
  - ▶ Export results

# Introduction

# Reproducible statistical analysis

Open principles applied to the way you conduct statistical data analysis:

- ▶ Make the process explicit and transparent
- ▶ Provide every input required to reproduce the analysis carried out and obtain the same results, as reported in the final document published

This typically involves three levels:

- ▶ **Data** used for the study
- ▶ **Code** created to perform the analysis
- ▶ **Platform** required to run the code

Being fully open on the three is not always possible (e.g. proprietary data/software), but that should be goal to which to get as close as possible.

*Getting halfway is better than not starting*

In this session we will focus on the last two: **code** and **platform**

# Reinhart & Rogoff

- ▶ In 2010, C. Reinhart and K. Rogoff put together a paper claiming to show how economic growth is seriously dampened once the ratio of debt to GDP goes above 90%
- ▶ The paper was very influential and became one of the most commonly cited ones to argue for austerity measures
- ▶ In 2013, **Thomas Herndon**, a PhD student at UMass, tried to replicate the results for a class assignment
- ▶ He could not, so finally he obtained from Reinhart the original (Excel) code and data only to find **results diverged** because of:
  - ▶ Selective exclusion of available data
  - ▶ Unconventional weighting of summary statistics
  - ▶ Coding errors
- ▶ The **replication** is posted online, together with the data and R code used for the paper

## Lessons:

- ▶ **No one is free from mistakes** (even Harvard top economists!)
- ▶ **Posting your data and code** but, if you don't, sharing them honestly upon request is a good second best
- ▶ **Replication** should be much more widespread
- ▶ ... you should not underestimate PhD students without a big name but with lots of time!

R



## R: what is it?

*R is a language and environment for statistical computing and graphics*

- ▶ **language & environment**
- ▶ **statistical computing**
- ▶ **graphics**

Characteristics:

- ▶ It is a Free implementation of the S language created by **Ross Ihaka** and **Robert Gentleman** in 1993
- ▶ **Cross-platform**: runs on many \*nix (included Linux) systems, Windows and MacOS.
- ▶ It is licensed under GPL, which makes it **free**. . .
  - ▶ ... as in **beer**
  - ▶ ... as in **speech**

# Why should I care about R?

- ▶ Philosophy behind the project
- ▶ Convenience (once you get ahead the learning curve)

Some people who care about R:

- ▶ Many top universities use R in teaching and research
- ▶ Google and Facebook
- ▶ New York Times

# The R Philosophy

*... Then sit back, relax, and enjoy being part of something big. . .*

[Tom Preston-Werner]

Being Free Software (“the users have the freedom to run, copy, distribute, study, change and improve the software”) has enhanced:

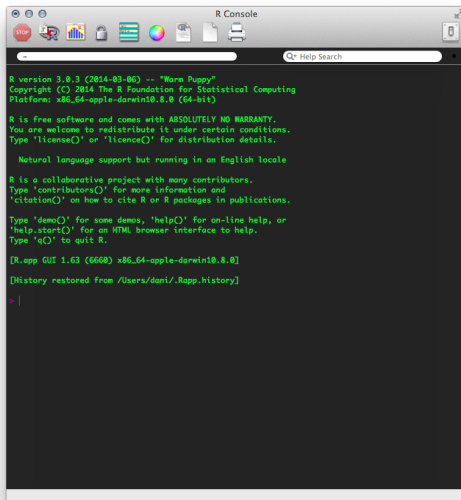
- ▶ **Worldwide community** of dedicated and enthusiastic users, contributors and developers that:
  - ▶ Lowers the entry barriers (mailing lists, blog posts, online tutorials, workshops. . .)
  - ▶ Continuously expands the capability and functionality
- ▶ Becoming an instrument for **democratization** of academic software and technology transfer
- ▶ Becoming the **lingua franca** in academia
- ▶ Facilitating reproducibility and Open Science

# R as free beer

- ▶ The price is right
  - ▶ Education
  - ▶ Installation across multiple machines
- ▶ The *beer selection* is wide (CRAN hosts 3,669 available packages as of March 10th. 2012)
  - ▶ Makes R a good one stop-shop and a good investment of your time to learn it
  - ▶ No market profitability constraints put it at the cutting edge (research sandbox)
- ▶ Linus' Law: *"given enough eyeballs, all bugs are shallow"*
  - ▶ More reliable and stable

# Ways to interact with R

## ► Interactive shell



```
R version 3.0.3 (2014-03-06) -- "Warm Puppy"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin10.0.0 (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.

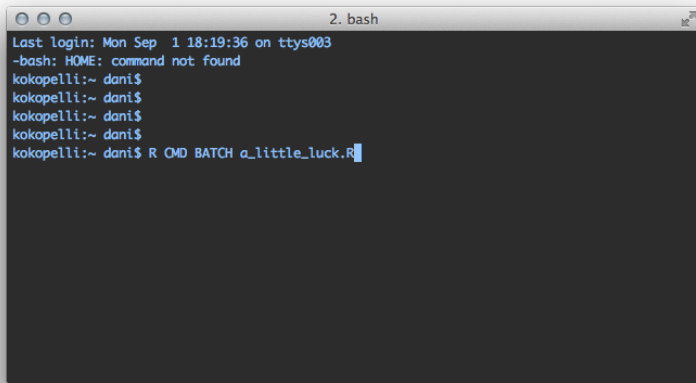
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.63 (6668) x86_64-apple-darwin10.0.0]
[History restored from /Users/dani/.Rapp.history]

> |
```

# Ways to interact with R

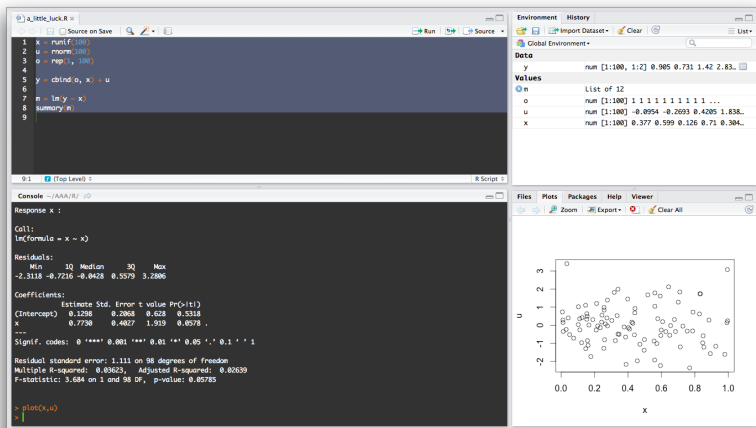
- ▶ Batch mode from the command line



```
2. bash
Last login: Mon Sep  1 18:19:36 on ttys003
-bash: HOME: command not found
kokopelli:~ dani$
kokopelli:~ dani$
kokopelli:~ dani$
kokopelli:~ dani$
kokopelli:~ dani$ R CMD BATCH a_little_luck.R
```

# Ways to interact with R

## ► IDEs (e.g. RStudio)



# R overview



# Packages

## Look for R info and packages

- ▶ Project website: `http://r-project.org`
- ▶ The Comprehensive R Archive Network (CRAN)
- ▶ The R-Journal (and JoSS)
- ▶ R bloggers
- ▶ Twitter: the `#rstats` hashtag
- ▶ Google (good luck on that)

## Install and load packages

- ▶ Windows and MacOS GUIs have installers
- ▶ Command line with `install.packages` function
- ▶ Command `library` (e.d. `library(maptools)`) to load the package `maptools`)

# Help

NEVER HAVE I FELT SO  
CLOSE TO ANOTHER SOUL  
AND YET SO HELPLESSLY ALONE  
AS WHEN I GOOGLE AN ERROR  
AND THERE'S ONE RESULT  
A THREAD BY SOMEONE  
WITH THE SAME PROBLEM  
AND NO ANSWER  
LAST POSTED TO IN 2003



# Help and documentation

- R built-in search capability

| Command                               | Function  |
|---------------------------------------|---|
| <code>?read.csv</code>                | Check local documentation for <code>read.csv</code> function                        |
| <code>spdep::moran.test</code>        | Check local documentation in package <code>spdep</code> for <code>moran.test</code> |
| <code>help("read.csv")</code>         | Check local documentation for <code>read.csv</code> function                        |
| <code>help.search("read.csv")</code>  | Search for “read.csv” in all help files   |
| <code>RSiteSearch("plot maps")</code> | Search for the term “plot maps” in the RSiteSearch website (requires connectivity)  |

- StackOverflow

# Reading data

Point to the folder

Native csv reading

# Exploring a data.frame

# Manipulating a data.frame

# Analyze data: regression

# Visualization



# Export results

More

## Additional resources