First Steps to Data Analysis in R

4 October 2023
EARNCon 2023

Ben Zipperer
Economic Policy Institute

bzipperer@epi.org
@benzipperer

Goals

This is a crash course in using R. You will learn

- To perform basic data analysis in R
- To update, replicate, and share your work by writing code in R
- Enough fundamentals to explore other R resources

Materials for this class

`https://economic.github.io/data_bootcamp/`

Big picture outline for today

1. R/RStudio basics
2. Analyze simple data
   - national wage percentiles, by race
3. Analyze complex data
   - CPS microdata
   - calculate demographic profile of low-wage workers in Virginia
4. Basic programming in R

1. R/RStudio basics: tasks

R is free, widely used software for data analysis.

Rstudio is software that makes it easy to use R.

Now we will learn

- the layout of R/Rstudio
- some very basic R commands and functions
- how to store results in R

1. R basics: review

- R is essentially a very fancy calculator
- R uses functions (commands)
- Functions
  - have a name
  - often need you to specify inputs (arguments) in parentheses
  - create an output (object)
  - can be nested
  - are described in help files: `?function`
- We store objects with assignment arrow: `<-`

2. Analyze simple data: Calculate Black-white wage ratio, last 10 years

Source data

- Data easily accessible from EPI: `https://www.epi.org/data`
- Provided to you as .csv file: `epi_wage_percentiles.csv`

Gameplan

- Load the data into R
- Calculate Black-white wage differences
- Export the results

2. Analyze simple data: review

Workflow: load data, manipulate it, and save output

`read_csv(”filename.csv”)` loads csv file

`select(data, column1, column2, ...)` keeps *column1, column2, ...*

`filter(data, condition)` keeps rows satisfying *condition*

`arrange(data, column1, column2, ...)` sorts rows according to *column1, column2, ...*

`mutate(data, column = ...)` change or create *column* according to the rule ...

`write_csv(”filename.csv”)` save resulting data as csv file

3. Analyze complex data: tasks

How many workers earn low hourly wages in Virginia?

- We will need worker-level data with wage and state information
- A good candidate for this is the Census / BLS Current Population Survey
    - easily accessible via EPI: `https://microdata.epi.org/`
    - 2022 CPS provided in Stata format: `epi_cpsorg_2022.dta.zip`
- Let's calculate the share of workers earning less than $15 / hour

3. Analyze complex data: review

`haven::read_dta("filename.dta")` loads Stata data file

`count(data, var1, var2, ...)` tabulates *var1, var2, ...*

`summarize(data, function)` provides summary statistic outputted by *function*

`mean(var)` and `weighted.mean(var, w = weight)` calculate means of *var*

4. Basic R programming: what and why?

- We just learned how to do data analysis in R *interactively*
- In general you should write and run R scripts
- An R script will
    - provide a fully documented record of your work
    - allow you to tweak or extend your analysis more easily
    - aid replication by others (and yourself!)

Review

Today we learned to

1. Load and use R/RStudio
2. Analyze simple data: national wage percentiles, by race
3. Analyze complex data: profile of low-wage workers in Virginia
4. Code in R
    - always write and run R scripts
    - add comments to document your work
    - write better R code with the pipe: **%>%**
    - use packages

How to learn more?

Later today: Accessing public data with R

Other resources

- Work through your own analysis
- Hadley Wickham & Garrett Grolemund, *R for Data Science*: `https://r4ds.had.co.nz/`
- Kieran Healy, *Data Visualization*: `https://socviz.co/`