Data Science for Physicians (DS4P)

Peter Higgins

2018-06-05

Contents

Chapter 1

Prerequisites

Thank you for giving this e-book a try. This is designed for physicians or others analyzing health data who are interested in pursuing this field using the R language. We will assume that:

- you have access to a computer
- that you have access to the internet
- that you can download the current version of R, and
- that you have downloaded a current version of Rstudio.

1.0.1 To Install R:

- Open an internet browser and go to https://www.r-project.org.
- Click the "download R" link in the middle of the page under "Getting Started."
- Select a CRAN location (a mirror site) and click the corresponding link.

- Click on the "Download R for Windows" link at the top of the page.

- Click on the "install R for the first time" link at the top of the page.
- Click "Download R for Windows" and save the executable file somewhere on your computer. Run the .exe file and follow the installation instructions.

- Now that R is installed, you need to download and install RStudio.

1.0.2 To Install RStudio:

- Go to http://www.rstudio.com and click on the "Download RStudio" button.
- Click on "Download RStudio Desktop."
- Click on the version recommended for your system (Windows, Mac, Linux), and save the downloaded file. Run the file and follow the installation instructions.

This is a book written in **RMarkdown**.

Each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading `#`.

To compile this example to PDF, you need XeLaTeX. It is recommended that you install the TinyTeX package (which includes XeLaTeX): https://yihui.name/tinytex/.

Chapter 2

Introduction

There are many books about Data Science. Why does the world need another one, particularly one targeting physicians?

- There is a lot of health care data
- There are a lot of interesting questions in health care
- There are particular and challenging issues in doing data analysis with PHI (Protected Health Information)

Syllabus: Data Science for Physicians (DS4P)

- Instructor: Peter Higgins, MD, PhD, MSc (CRDSA), Professor of Internal Medicine
- Office Hours: MSRB One 6510
- In-person class time
  - MSRB One 6510, Thursday evenings 6:30-8:30 PM

2.0.1 Course Description and Objectives

2.0.1.1 Description

A practical introduction to data collection and security, data cleaning, statistical methods and computational tools needed to make sense of data, and methods for reporting and sharing your findings. This course is not a traditional introductory statistics courses in that computing plays a more central role than mathematics and a higher emphasis is placed on "thinking with data." Topics include

- secure HIPPA-compliant data collection
- data cleaning and validation
- data visualization
- data wrangling
- confidence intervals
- hypothesis testing, and
- regression The course has no mathematics or computer science prerequisites.

2.0.1.2 Objectives

1. Have students engage in the data/science research pipeline in as faithful a manner as possible while maintaining a level suitable for novices.
2. Foster a conceptual understanding of statistical topics and methods using real clinical data whenever possible, and simulation/resampling to support teaching concepts of inference.
3. Use a flipped classroom model by incorporating online learning for new concepts, with limited face-to-face time for real-time problem-solving
4. Introduce best practices for reproducible research and collaboration.
5. Develop statistical literacy by, among other ways, tying in the curriculum to actual clinical data, demonstrating the importance statistics and computing plays in advancing medicine

2.0.1.3 Topics

Roughly speaking we will cover the following topics (a more detailed outline is found below:

1. Introduction and Tools (R, RStudio, and R Markdown)
2. Data Import, and Handling
3. Data Collection
4. Checking, Validating, And Exploring your Data
5. Data Types
6. Data Wrangling with Tidyr and Dplyr
7. Graphic Summaries for a Single Variable – ggplot package
8. Descriptive Data for a Single Variable
9. Graphic Summaries for Two or More Variables – ggplot2
10. Descriptive Data for Two or More Variables
11. Presenting your Results in a report with RMarkdown
12. Statistical inference
13. Study Design
14. Sample Size and Power
15. Sources of Bias
16. Study Types
17. One variable, single group
18. One variable, two groups
19. Multiple groups
20. Linear Regression
21. Reporting results interactively with Shiny
22. Logistic Regression
23. Meta-Analysis

2.0.1.4 Learning Resources

- E-Textbooks: Open Intro Statistics, at www.openintro.org
- E-Books on R These are at different levels:

Level: Absolute Beginner Textbook: R Basics
Goal: Set up R and RStudio on a laptop, introduce the concept of an IDE
Link:

Level: New to R & Statistics
Textbook: Modern Dive Goal: Learn basics of Data Management and visualization, introduction to hypothesis testing and statistical modeling
Link:

Level: Comfortable with R
Textbook: Hands-On R Programming
Goal:
Link:

Level: Ready to Understand More
Textbook: R for Data Science
Link:

- Software:

  - Local laptop/desktop free open-source version of R and RStudio
  - Cloud-based RStudio Server, which you can access in your browser via: Note if you are off-campus you must first log into the UM VPN.

- Online:

  - DataCamp. A brower based interactive tool for learning R through short, focused courses, each 3-4 hours long.
  - RStudio. Website with many resources for learning about the RStudio IDE and the tidyverse.
  - r-cookbook – an often useful website with concrete examples of how to use R packages
  - Stack Overflow and Google. Remarkably helpful to search for explanations of error messages, or explanations of problems that someone else has probably also experienced. For using Google, search for any topic or your error message and add "in R"
  - package vignettes – variable quality, but when well done, can be extremely helpful examples of how to use the functions in each package
  - R twitter – follow Rbloggers, #rstats

2.0.1.5 Evaluation

This course is entirely voluntary. I hope that you will learn valuable skills that will advance your research career. I would like you to progress to using these skills on your own data as quickly as possible, as this will greatly help you reinforce your new skills. There are no grades and no formal evaluations. You can, however, earn certificates on DataCamp for completing courses.

2.0.1.6 Task Goals

1. Learn concepts through Data Camp

a. Multiple short courses to correspond with each unit

2. Test yourself with assignments in ModernDive

a. Chapters corresponding to each unit

3. Three Challenges

a. Clean data and perform descriptive data analysis on the biofire dataset
b. Clean data and model outcomes in the health satisfaction dataset, producing a final report

c. Use logistic regression to model dichotomous outcomes and produce a Shiny app to allow users to make predictions for future patients

4. Final Project There will be a final capstone project. This is an opportunity for you to use your statistics and data science skills developed during the challenges and perform your own start-to-finish data analysis project. The project will involving you addressing a scientific question by choosing a data set (or preferably, using one of your own), performing an analysis using the concepts and tools we have covered in this course, and writing a report. This can be done solo or with a partner.

2.0.1.7 Learning Goals

1. Recognize the importance of data collection, identify limitations in data collection methods, and determine how they affect the generalizability of your findings
2. Use statistical software (R) to summarize data numerically and visually, and to perform data analysis.
3. Have a conceptual understanding of statistical inference.
4. Apply estimation and testing methods to analyze single variables or the relationship between two variables in order to understand data relationships and make data-based conclusions.
5. Model numerical response variables and dichotomous response variables using a single explanatory variable or multiple explanatory variables in order to investigate relationships between variables.
6. Interpret results correctly, effectively, and in context without relying on statistical jargon.
7. Critique data-based claims and evaluate data-based decisions.

2.0.1.8 Tips for success

1. Read materials for each unit
2. Do Data Camp courses for each unit – usually around 1 chapter (1 hour) per day.
3. Do Data Camp daily practice on any day that you don’t have time to do a full chapter
4. At end of each course, review material, take notes, copy/reproduce/save code on your laptop
5. Try new skills on your own data, or on one of the open data sets
6. Use RStudio and DataCamp Cheat sheets
7. Annotate your code to help ‘future you’ understand it.
8. Save and reuse your code for future projects

2.0.1.9 Expected work load

This course is entirely voluntary. It is expected that you have lots of clinical and/or research work to keep up with, along with the occasional call or night rotation. This is an investment in future skills to help your career. I recommend that you try to do up to one hour a day on most days, and on days when that is not realistic, to just do the 10 minutes of daily practice on DataCamp to keep the information fresh in your mind.
Other learning resources:

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 2. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter **??**.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```r
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 2.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 2.1.

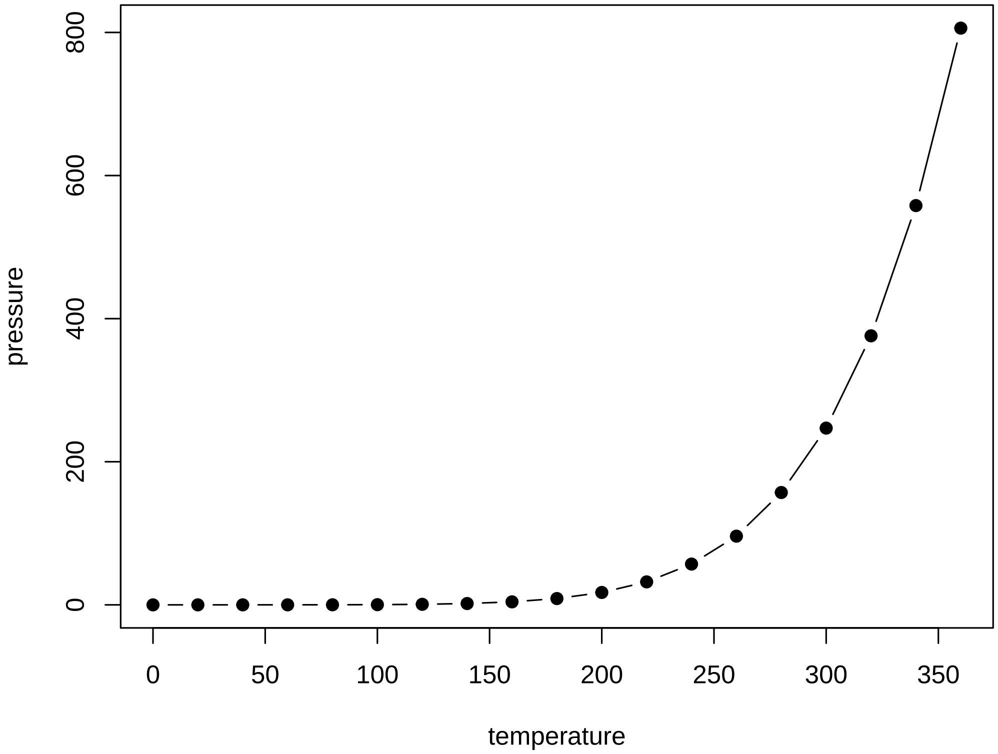

Figure 2.1: Here is a nice figure!

```
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2018) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

Table 2.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

Chapter 3

Starting Out with R and RStudio

3.0.0.1 Introduction and Tools (R, RStudio, and R Markdown)

3.0.0.2 Install R on your computer

3.0.0.3 Install RStudio

3.0.0.4 Access DataCamp online

3.0.0.5 DataCamp for RStudio IDE (Part 1)

3.0.0.6 DataCamp for RStudio IDE (Part 2)

3.0.0.7 Access RStudio cloud

3.0.0.8 R basics E-book

Use the e-book Rbasics by Chester Ismay

https://ismayc.github.io/rbasics-book/

3.0.0.9 RStudio tips document

Chapter 4

Importing Your Data

4.0.0.1 Lots of options – learn rio

4.0.0.2 Install rio

4.0.0.3 Practice loading data from multiple file types

4.0.0.4 Practice saving as csv, rds, xls, xlsx

4.0.0.5 DataCamp Courses on Import (Part 1)

4.0.0.6 Data Camp Course on Import (Part 2)

4.0.0.7 Modern Dive Chapter 1

4.0.0.8 Modern Dive Chapter 2

4.0.0.9 Chapter Challenges

Chapter 5

Data Collection

Some *significant* applications are demonstrated in this chapter.

5.0.0.1 Best practices for data in spreadsheets

https://peerj.com/preprints/3183/

5.0.0.2 Google Forms and GoogleSheets

5.0.0.3 SurveyMonkey data

5.0.0.4 PHI data – REDCap and redcapr package

5.0.0.5 Issues with PHI – laptops, memory sticks, cloud, github

5.0.0.6 PHI solutions - private github, M+Box, deidentifying PHI with charlatan package in R, link to PHI that is secure

5.0.0.7 Other sources of data

5.0.0.7.1 Surveys – REDCap and SurveyMonkey

5.0.0.7.2 Data Direct

5.0.0.7.3 Data Warehouse

Chapter 6

Checking, Validating, And Exploring your Data

6.0.0.1 Cleaning – names with janitor package to snake_case

6.0.0.1.1 A few words about tidyverse style

6.0.0.2 Finding Missing data – naniar and visdat packages

6.0.0.3 Validating data – validate package

6.0.0.4 Evaluating – str, glimpse

6.0.0.5 Exploring- skimr package

6.0.0.6 Histograms

6.0.0.7 Correlations – ggally extension of ggplot2, and corrr package

Chapter 7

Data Types

7.0.0.1 Numeric - Integer, double

7.0.0.2 Strings with the stringr package

7.0.0.2.1 Rebus package and regex

7.0.0.2.2 DataCamp strings course

7.0.0.3 Factors with the forcats package

7.0.0.3.1 https://peerj.com/preprints/3163/

7.0.0.4 Dates with the lubridate package

7.0.0.4.1 DataCamp dates and times course

Chapter 8

Data Wrangling with Tidyr and Dplyr

8.0.0.1 What is Tidy Data?

8.0.0.2 DataCamp Tidyr course

8.0.0.2.1 Try Tidying New Data - Challenges

8.0.0.3 What is Data Wrangling?

8.0.0.4 DataCamp Dplyr wrangling course

8.0.0.4.1 Try Wrangling New Data - Challenges

8.0.0.4.2 General Principles of Tidying And Wrangling

https://peerj.com/preprints/3180/

8.0.0.5 DataCamp Dplyr joins course

8.0.0.5.1 Try Joining New Data - Challenges

Chapter 9

Graphic Summaries for a Single Variable with ggplot

9.0.0.1 DataCamp ggplot 1 course

9.0.0.2 Histograms

9.0.0.2.1 Histogram Challenges

9.0.0.3 Boxplot

9.0.0.3.1 Boxplot Challenges

9.0.0.4 Violin plot

9.0.0.4.1 Violin Plot Challenges

Chapter 10

Graphic Summaries for a Single Variable with ggplot

10.0.0.1 DataCamp ggplot 1 course

10.0.0.2 Histograms

10.0.0.2.1 Histogram Challenges

10.0.0.3 Boxplot

10.0.0.3.1 Boxplot Challenges

10.0.0.4 Violin plot

10.0.0.4.1 Violin Plot Challenges

Chapter 11

Graphic Summaries for Two or More Variables – ggplot2

11.0.0.1 Data Camp ggplot course 2

11.0.0.2 Scatter plots

11.0.0.3 Mosaic plots

11.0.0.4 Corrr package

11.0.0.5 Correlation matrix

Chapter 12

Descriptive Data for Two or More Variables

12.0.0.1 Table

12.0.0.2 Janitor crosstab tabyl

Chapter 13

Presenting your Results in a report with RMarkdown

13.0.0.1 DataCamp Rmarkdown Course

13.0.0.2 Practice with data from Chapter 8 - HTML

13.0.0.3 Practice with data from Chapter 9 - Word

13.0.0.4 Practice with data from Chapter 10 - PPT

Chapter 14

Reproducibility in Your Research

14.0.0.1 Collaborating with Past You and Future You

14.0.0.1.1 General references

14.0.0.1.1.1 https://peerj.com/preprints/3192/

14.0.0.1.1.2 https://peerj.com/preprints/3139/

14.0.0.2 DataCamp GitHub Course

14.0.0.3 The problem of versions and updated packages

14.0.0.3.1 Solutions

14.0.0.3.1.1 Packrat

14.0.0.3.1.2 Microsoft R checkpoinT

14.0.0.3.1.3 Rocker (docker) containers

14.0.0.4 R Projects

14.0.0.5 RStudio on Projects

14.0.0.5.1 Multiple scripts and organization of projects

14.0.0.5.2 Version control

14.0.0.5.2.1 https://peerj.com/preprints/3159/

14.0.0.6 Linear and branching projects, and use of the drake package

Chapter 15

Statistical inference

15.0.0.1 Concepts

15.0.0.2 DataCamp Stats courses

15.0.0.3 ModernDive chapters

15.0.0.4 Infer package and practice

Chapter 16

Study Design

16.0.0.1 Hypotheses – null and alternative

16.0.0.2 Specific and testable

16.0.0.3 Inclusion and exclusion

16.0.0.4 How many eligible patients – DataDirect query

16.0.0.5 Be conservative – most won't enroll

Chapter 17

Sample Size and Power

17.0.0.1 Estimating measures in each group

17.0.0.2 Estimating effect size

17.0.0.3 Estimating range and SD

17.0.0.4 Power calc for continuous

17.0.0.5 Power calc for proportion

Chapter 18

Sources of Bias

18.0.0.1 Generalizability of studies

18.0.0.2 Bias in design

18.0.0.3 Recall Bias

18.0.0.4 Time bias

18.0.0.5 Ascertainment bias

18.0.0.6 Assuming linearity on a scale

Chapter 19

Study Designs

19.0.0.1 Cross sectional

19.0.0.2 Cohort

19.0.0.3 Retrospective

19.0.0.4 Rarely prospective – registries, case series more likely

19.0.0.5 All Associations

19.0.0.6 Causal Inference requires randomization

Chapter 20

One variable, single group

20.0.0.1 Continuous value – t test

20.0.0.2 Challenges for single continuous outcome

20.0.0.3 Estimate single Proportion

20.0.0.4 Challenges for single proportion

Chapter 21

One variable, two groups

21.0.0.1 Variable 1 is greater in group A vs group B

21.0.0.1.1 Test for skew – if not, T test

21.0.0.1.1.1 Challenges

21.0.0.1.2 If yes, Wilcoxon, non parametric

21.0.0.1.2.1 Challenges

21.0.0.2 Variable 2 proportion is greater in group A vs group B

21.0.0.2.1 If no rare cells, chi square

21.0.0.2.1.1 Challenges

21.0.0.2.2 If rare cells, Fischer exact test

21.0.0.2.2.1 Challenges

Chapter 22

One variable, Multiple groups

22.0.0.1 DataCamp

22.0.0.2 Fun with ANOVA

22.0.0.2.1 Challenges

22.0.0.3 Consider Regression if complicated

Chapter 23

Linear Regression

23.0.1 DataCamp Course

23.0.1.1 Single outcome, multiple possible predictors

23.0.1.1.1 Challenges

23.0.1.2 One predictor at a time – multiple univariate models – modelr and broom packages

23.0.1.2.1 Challenges

23.0.1.3 Choosing predictors for multivariate modeling – testing, dealing with collinearity

23.0.1.3.1 Challenges

23.0.1.4 Multivariate modeling

23.0.1.4.1 Challenges

23.0.1.5 Model fit checking

23.0.1.5.1 Challenges

23.0.1.6 presenting model results with RMarkdown

23.0.1.6.1 Challenges

Chapter 24

Sharing Models with Shiny

24.0.0.1 DataCamp Courses on Shiny

24.0.0.2 Practice with model from 20

Chapter 25

Logistic Regression

25.0.0.1 Concepts and OR

25.0.0.2 Modeling

25.0.0.3 Estimating GOF

25.0.0.4 AuROC

25.0.0.5 Sens Spec PPV NPV

25.0.0.6 Confusion Matrix

25.0.0.7 Present results with broom, RMarkdown

25.0.0.8 Make user-friendly model predictions with Shiny

Chapter 26

Meta-Analysis

26.0.0.1 Data search

26.0.0.2 Data collection

26.0.0.3 Data Exclusion

26.0.0.4 Data extraction and checking

26.0.0.5 Using Metafor package

26.0.0.6 Making Figures

26.0.0.7 Writing up results in RMarkdown

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr.* Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2018). *bookdown: Authoring Books and Technical Documents with R Markdown.* R package version 0.7.