Stat 850: Computing Tools for Statisticians

Susan Vanderplas

2020-05-12

Contents

Introduction

Chapter 1

Tools for Statistical Computing

This course requires that you have the following software installed on your machine:

- git
- SAS 9.4 (or later)
- R (3.5 or higher)
- RStudio
- LaTeX

You will also need to sign up for a GitHub account

This chapter will provide instructions for how to install and configure these programs (or how to obtain them), but first, it may be useful to get a "big picture" overview of how the course is structured

1.1 Course Outline

The goal of this class is to expose you to basic computing skills in R and SAS, which are two of the more common languages for statistical computing (python is the 3rd most common, and is particularly popular in data science and machine learning, but will not be explicitly taught in this class.)

SAS is another extensively used statistical programming language. ...

1.2 R and RStudio

R is a statistical computing language which originated as an open-source clone of Bell labs S computing language. S was inspired by Scheme, but also has features which are similar to Lisp. It is a scripting language (you don't have to compile the code before it runs) and is natively accessed using a command-line prompt. One feature of R that is relatively unique is that it uses vector-based math, which means that mathematical operations on vectors occur for the entire vector without having to use loops to iterate through the vector line-by-line. R is optimized for working on data: unlike more general-purpose programming languages such as Python, R was built with the idea of facilitating data analysis. As a result, data structures in R tend to be more natural for statistical work. From a computer science perspective, though, R seems like an extremely odd language because the design choices that make data analysis easier are unconventional for more general-purpose languages.

RStudio is an integrated development environment for R. Basically, it adds a pretty graphical layer on top of R, providing an easy way to develop R code, evaluate that code, and keep track of all of the variables which are available in the computing environment. RStudio contains integrations which provide syntax highlighting, code folding, basic checks (missing parentheses, etc.), and many other features. RStudio was designed around the idea of making R easier to use and making it easy to develop statistical software reproducibly. RStudio (the company) is responsible for adding many features to the R ecosystem which facilitate running statistical analyses and presenting the results in user-friendly ways.

1.2.1 Getting Set up: R

1.2.2 Getting Set up: RStudio

1.2.3 Exploring RStudio

1.3 SAS

1.4 LaTeX

LaTeX purpose

Overleaf vs. TexMaker and other GUIs

For now, LaTeX is going to sit there under the hood and be used by other programs

1.5 Version Control with Git

Explanation of why version control matters (phD comic)

1.5.1 Getting set up: git

1.5.2 Getting set up: github

1.5.3 Git and Github

Explain the difference between the two

Chapter 2

Introduction to Statistical Programming

2.1 Programming Concepts: Variables, Structures, and more

2.1.1 Variable types

2.1.2 Data structures

2.1.3 Control structures

2.2 Overgrown Calculators

R and SAS for basic numerical operations

Matrix algebra

2.3 Basic Programming in R

2.4 Basic Programming in SAS

Chapter 3

Organization: Packages, Functions, Scripts, and Documents

3.1 Reproducibility

Why reproducibility is important in science and statistics

Why reproducibility is convenient

3.2 Markdown

3.3 Functions and Procs

3.4 Scripts

3.5 Packages

3.6 Comparisons - SAS and R

Chapter 4

External Data

4.1 External Data Formats

- txt
- tsv
- csv
- fixed width
- excel/spreadsheet
- databases
- structured data (xml, json)

4.2 Importing Tabular Data into R

4.3 Importing Tabular Data into SAS

4.4 Exploratory Data Analysis

- tables
- summary statistics
- basic plots?
- unique values

Chapter 5

Manipulating Data

5.1 Filter

5.2 Select

5.3 Group By

5.4 Summarize

Chapter 6

Transforming Data

6.1 Pivot operations

6.1.1 Wider

6.1.2 Longer

6.2 Separating one variable into many

Chapter 7

Introduction to Data Analysis

7.1 Models in SAS

7.2 Models in R

7.2.1 Standard Models

7.2.2 Tidymodels

Chapter 8

Introduction to Data Visualization

8.1 Creating Charts in SAS

8.2 Creating Charts in R (base)

8.3 The Grammar of Graphics

Introduce grammar of graphics as implemented in ggplot2

Chapter 9

Graphical Communication

9.1 something

Chapter 10

Simulation and Reproducibility

10.1 Pseudorandom Number Generation

10.2 Random Number Generation and Reproducibility

10.2.1 Seeds

10.2.2 Rmarkdown Reproducibility

Chapter 11

Principles of Debugging

11.1 Ideal Function Design

11.2 Debugging Tools in R

11.3 Debugging Tools in SAS

11.4 Minimal Working Examples

11.5 Researching Error Messages

Chapter 12

Dynamic Presentations

12.1 Slides

12.1.1 Beamer (LaTeX) and knitr

12.1.2 xaringan

12.2 Posters

12.2.1 Beamer (LaTeX) and knitr

12.2.2 Pagedown

Chapter 13

SAS in bookdown

```
proc printto log="saslog.log" new;
proc means data=sashelp.class;
run;
```

```
<div class="branch">
<a name="IDX"></a>
<div>
<div align="center">
<!--BEGINTABLE--><table class="table" cellspacing="0" cellpadding="7" rules="groups" frame="hside
<colgroup>
<col>
</colgroup>
<colgroup>
<col>
<col>
<col>
<col>
<col>
</colgroup>
<thead>
<tr>
<th class="l b header" scope="col">Variable</th>
<th class="r b header" scope="col">N</th>
<th class="r b header" scope="col">Mean</th>
<th class="r b header" scope="col">Std Dev</th>
<th class="r b header" scope="col">Minimum</th>
<th class="r b header" scope="col">Maximum</th>
</tr>
</thead>
<tbody>
```

```
<tr>
<th class="l stacked_cell data"><table width="100%" border="0" cellpadding="7" cellspac
<tr>
<th class="l data top_stacked_value">Age</th>
</tr>
<tr>
<th class="l data middle_stacked_value">Height</th>
</tr>
<tr>
<th class="l data bottom_stacked_value">Weight</th>
</tr>
</table></th>
<td class="r stacked_cell data"><table width="100%" border="0" cellpadding="7" cellspac
<tr>
<td class="r data top_stacked_value">19</td>
</tr>
<tr>
<td class="r data middle_stacked_value">19</td>
</tr>
<tr>
<td class="r data bottom_stacked_value">19</td>
</tr>
</table></td>
<td class="r stacked_cell data"><table width="100%" border="0" cellpadding="7" cellspac
<tr>
<td class="r data top_stacked_value">13.3157895</td>
</tr>
<tr>
<td class="r data middle_stacked_value">62.3368421</td>
</tr>
<tr>
<td class="r data bottom_stacked_value">100.0263158</td>
</tr>
</table></td>
<td class="r stacked_cell data"><table width="100%" border="0" cellpadding="7" cellspac
<tr>
<td class="r data top_stacked_value">1.4926722</td>
</tr>
<tr>
<td class="r data middle_stacked_value">5.1270752</td>
</tr>
<tr>
<td class="r data bottom_stacked_value">22.7739335</td>
</tr>
</table></td>
<td class="r stacked_cell data"><table width="100%" border="0" cellpadding="7" cellspac
```

```
<tr>
<td class="r data top_stacked_value">11.0000000</td>
</tr>
<tr>
<td class="r data middle_stacked_value">51.3000000</td>
</tr>
<tr>
<td class="r data bottom_stacked_value">50.5000000</td>
</tr>
</table></td>
<td class="r stacked_cell data"><table width="100%" border="0" cellpadding="7" cellspacing="0">
<tr>
<td class="r data top_stacked_value">16.0000000</td>
</tr>
<tr>
<td class="r data middle_stacked_value">72.0000000</td>
</tr>
<tr>
<td class="r data bottom_stacked_value">150.0000000</td>
</tr>
</table></td>
</tr>
</tbody>
</table>
<!--ENDTABLE--></div>
</div>
<br>
</div>
```

```
cat(readLines("saslog.log"), sep="\n")
```

```
NOTE: PROCEDURE PRINTTO used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds


7          proc means data=sashelp.class;
8          run;

NOTE: There were 19 observations read from the data set SASHELP.CLASS.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.00 seconds
      cpu time            0.01 seconds


9
```

```
# Do not forget to remove the log file when you are done!
unlink("saslog.log")
```