

Atomic data objects in R

SDIV

1/16/23

Table of contents

Preface	3
Introduction	4
The tidyverse	4
 I Background	 6
1 R programming fundamentals for Data Science	8
1.0.1 Entering Input: the assignment operator	8
1.1 Best Coding Practices for R	8
1.1.1 What we mean when say “better coding practice”	8
1.1.2 Folder Structure	9
1.1.3 Code Structure	9
1.1.4 Sections	9
1.1.5 Structural Composition	9
1.1.6 Indentation	9
1.1.7 Styling	9
1.1.8 Final Comments	9
 2 Data visualization with ggplot2 and friends	 10
 II The whole game of statistical Inference	 11
3 Statistical Inference with resampling: Bootstrap and Jackknife.	13
3.1 Likelihood inference.	13
3.2 Variance analysis.	13
3.3 ROC Curves	13
 4 Linear Regression	 14
4.1 Linear Regression	14
4.2 Multiple linear regression and generalized linear regression	14
 5 Summary	 15
 References	 16

Preface

- Who I am. I am Saul Diaz Infante Velasco. I just starting as assistant professor at the Data Science graduate program of Universidad de Sonora at Hermosillo Mexico. My Background is related with numerical analysis and stochastic models. I'm are a enthusiastic of this treading topic called Data-Science, but perhaps at the moment I only have just intuition about what really it is. However, I have been programming almost 20 years an moved from old programming langues as FORTRAN, Pascal, Basic, Cobol, C, C++ to thenew well established treading development workflows like R, Python and Julia. This is my firs attempt in R.
- What the book is about.
- When I writing this book.
- Why I write this book.
- Where I wrote this book.

Introduction

The focus of this course is into the programming and basic techniques for inference that are usually applied in data science. We start by reviewing and enforcing programming skills. Then we will use the database of entomological data practice and build the required bases for more structured tools like bootstrap or Jackknife cuts.

Figure 1 further explores the impact of temperature on ozone level.

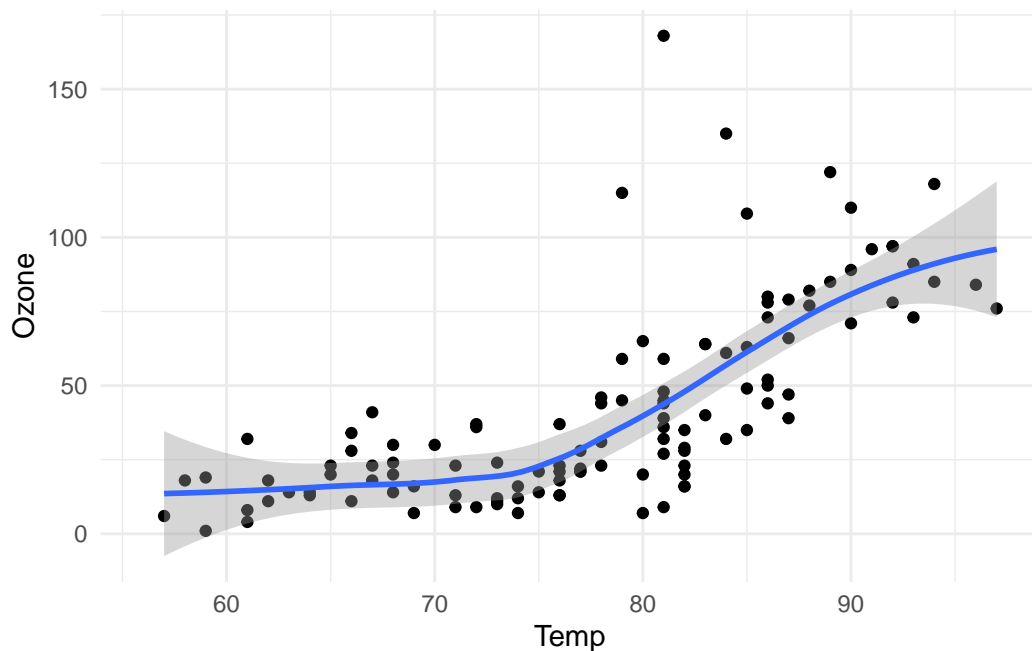


Figure 1: Temperature and ozone level.

The tidyverse

We need to install a R package. The majority of the packages that we will use are part of the so-called tidyverse package. The packages in the tidyverse share a common philosophy of data and R programming, and are designed to work together naturally.

You can install the complete tidyverse with the line of code:

then we can use it by loading in the preamble section with

```
-- Attaching packages ----- tidyverse 1.3.2 --
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.3.0      v stringr 1.5.0
v readr   2.1.3      v forcats 0.5.2
v purrr   1.0.1
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
```

see <https://www.tidyverse.org/> documentation.

Part I

Background

We dedicate this part to overview the basics to program in R. The aim of this part is building the basis for Machine learning, namely **data visualization**, **data manipulation** and the **good coding practices** to type script of industrial production quality.

1 R programming fundamentals for Data Science

1.0.1 Entering Input: the assignment operator

The thing that we type on the R console prompt are expressions. The first expression we discuss here is the assignment operator, please watch the following video https://www.youtube.com/watch?v=vGY5i_J2c-c&t=283s

At the R console, any executable typed text that we put a side of the prompt are called expressions. We start by the `<-` symbol is the assignment operator.

```
[1] 0
```

```
[1] 0
```

```
[1] "what's up"
```

The `[1]` shown in the output indicates that `x` is a vector and `0` is the element at position with index 1.

1.1 Best Coding Practices for R

1.1.1 What we mean when say “better coding practice”

R programmers have a bad reputation writing bad code. Perhaps the main reason is that the people who write much of the package are not programmers but scientific from other areas. Sometimes we overestimate crucial aspects from a programming standpoint. As R programmers we overcome to write the code for production. Mostly we write scripts and when we deploy it the same when we just wrap it in a function and perhaps a package. It is common to face poorly written code—**columns were referred by numbers, functions were dependent upon global environment variables, 50+ lines functions without arguments and with over-sized lines code 100 characters or more, not indentation, poor naming, conventions etc.,...**

We strongly encourage to use a style. Yea I know, there is not a unique way to do it, but the philosophy is to follow a consistent style. With respect to this regard made yourself a favor and read this great book for R

<https://bookdown.org/content/d1e53ac9-28ce-472f-bc2c-f499f18264a3/>

1.1.2 Folder Structure

1.1.3 Code Structure

1.1.4 Sections

1.1.5 Structural Composition

1.1.6 Indentation

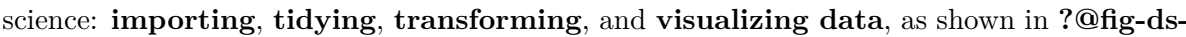
1.1.7 Styling

1.1.8 Final Comments

2 Data visualization with ggplot2 and friends

Part II

The whole game of statistical Inference

Our goal in this part of the book is to give you a rapid overview of the main tools of data science: **importing**, **tidying**, **transforming**, and **visualizing data**, as shown in  **fig-ds-whole-game**. We want to show you the “whole game” of data science giving you just enough of all the major pieces so that you can tackle real, if simple, data sets. The later parts of the book, will hit each of these topics in more depth, increasing the range of data science challenges that you can tackle.

3 Statistical Inference with resampling: Bootstrap and Jackknife.

3.1 Likelihood inference.

3.2 Variance analysis.

3.3 ROC Curves

4 Linear Regression

4.1 Linear Regression

4.2 Multiple linear regression and generalized linear regression

5 Summary

In summary, this book has no content whatsoever.

[1] 2

References

- [1] T. Hastie, R. Tibshirani, J. Friedman, [The elements of statistical learning](#), Second, Springer, New York, 2009.
- [2] W.J. Krzanowski, D.J. Hand, [ROC curves for continuous data](#), CRC Press, Boca Raton, FL, 2009.
- [3] R. Martin, [A statistical inference course based on p-values](#), The American Statistician. 71 (2017) 128–136.
- [4] P. McCullagh, J.A. Nelder, [Generalized linear models](#), Chapman & Hall, London, 1989.
- [5] B. Ratner, [Statistical and machine-learning data mining:: Techniques for better predictive modeling and analysis of big data](#), third edition, CRC Press, 2017.
- [6] D.A. Sprott, Statistical inference in science, Springer-Verlag, New York, 2000.