Chapter 1: Data Analysis and Data Science

Mark Andrews

Contents

Introduction

This book is about statistical data analysis of real world data using modern tools. It is aimed at those who are currently engaged in, or planning to be engaged in, analysis of statistical data of the kind that might arise at or beyond PhD level scientific research, especially in the social sciences. The data in these fields is complex. There are many variables and complex relationship between them. Analyzing this data almost always requires data wrangling, exploration, and visualization. Above all, it involves modelling the data using flexible probabilistic models. These models are then used to reason and make predictions about the scientific phenomenon being studied. This book aims to address all of these topics. The term we use for these topics and their corresponding methods and tools is *data science*.

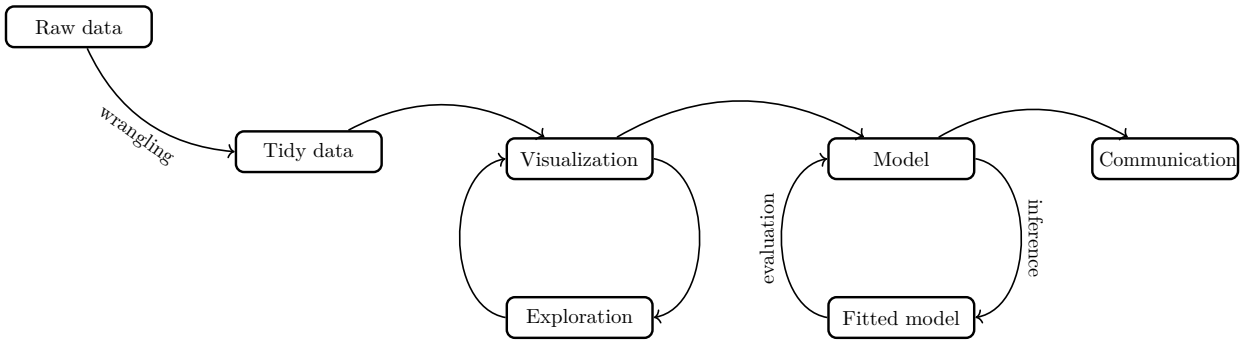

Figure 1: The data science workflow. Raw data is usually messy and not yet amenable to analysis of any kind. *Data wrangling* takes the raw data and transforms it into a new *tidy* format. This data is then explored and visualized in an iterative manner, which may also include some further wrangling. This eventually leads to probabilistic modelling, which itself involves an iterative process of statistical inference and model evaluation. Finally, we communicate our results in articles, presentations, webpages, etc.

As we use the term throughout this book, *data science* is a set of interrelated computational or mathematical methods and tools that are used in the general data analysis workflow that we outline in Figure 1. This workflow begins with data in its nascent and raw form. Raw data is usually impossible or extremely difficult to work with, even casually or informally. The process of transforming the data so that it amenable to further analysis is *data wrangling*, and the resulting data sets are said to be *tidy*. This data can then be explored and visualized. We view data exploration and data visualization as ultimately accomplishing the same thing. One usually involves quantitative descriptive analysis, while the other involves graphical analysis, but both aim to discover potentially interesting patterns and behaviours in the data. The exploratory analysis stage then leads us to posit a tentative probabilistic model of the data. Put more precisely, it leads us to posit a tentative probabilistic model of the phenomenon that generated the data. Inevitably, this model involves unknown variables that must be inferred using statistical inference. This leads to a fitted model, which may be then evaluated and possibly extended and modified, thus leading to further inference. Eventually, we communicate our results in reports, presentations, webpages, etc.

Each of the stages of this data science workflow involves involves computational and mathematical concepts and methods. In fact, it is this combination of the computational and the mathematical or statistical that is a defining feature or key characteristic of data science as we conceive of it. Without using computers, and thus performing any stages of the workflow manually in some manner, then only practically trivial types of analysis could be accomplished, and even then the analysis would be labourious and error prone. By contrast, the more proficient we are with the relevant computational tools, the more efficient and sophisticated our analyses can be. In this sense, computing skills, specifically reading and writing code, are integral and vital parts of modern data analysis. These can not be generally sidestepped or avoided by using graphical user interfaces to statistics programs. While programs like these are sometimes suitable for novices or for casual use, they are profoundly limited and inefficient in comparison to writing code in a high level programming language.

In addition to computing tools, many of the stages of the data science workflow involve mathematical and statistical concepts and methods. This is especially true of the statistical modelling stage, which requires a proper understanding of mathematical and probabilistic models, and related topics such as statistical inference. Simply being able to perform a statistical analysis computationally, accompanied by a vague and impressionistic understanding of what the analysis is doing and why it is doing it, will not generally be sufficient. Without a deeper and theoretical understanding of probabilistic models, statistical inference, and related concepts, we will not be able to make principled and informed choices concerning which models to use for any given problem. Nor would we be able to understand the meaning of the results of the inference, and we would be limited or mistaken in the practical and scientific conclusions that we make when we use these models for explanation or prediction. Moreover, statistical models of the kind that we cover in this book should not be seen as list of independent tools in a big toolbox, each one designed for a different task or application, and each with their own rules and principles. Rather, more generally, we should view statistical modelling as a systematic framework, or even a language, for building mathematical models of scientific phenomenon using observed data. While we may talk about normal linear models, or zero inflated Poisson models, etc., these are just examples of the infinitely many models that we can build to model the scientific problem at hand. Being aware of statistical modelling as a flexible and systematic framework that is based on pragmatic and theoretical principles allows us to more competently and confidently perform statistical analysis, and also greatly increases the range and scope of the analyses that are readily available to us.

What is data science?

Even if we accept the nature and the value of the data analysis workflow that we've just outlined, it is reasonable to ask whether it should properly be called data science. Is this not just using a new word, even a buzzword, in place of much more established terms like statistics or statistical data analysis? We are using the term data science rather than statistics per se or some variant thereof because data analysis as we've outlined it arguably goes beyond the usual focus of statistics, at least as it is traditionally understood. Mathematical statistics as a scientific or mathematical discipline has focused largely on the statistical modelling component of the program we outlined above. As we've hopefully made clear, this component is of profound importance,

and in fact we would argue that it is the single most important part and even ultimate goal of data analysis. Nonetheless, in practice, data wrangling alone occupies far more of our time and effort in any analysis, and exploration and visualization should be seen as necessary precursors to, and even continuous with, the statistical modelling itself. Likewise, traditional statistics often marginalizes the practical matter of computing tools. In statistics textbooks, even excellent ones, for example, code examples may not be provided for all analyses, and the code may not be integrated tightly with the coverage of the statistical methods. In this sense, traditional statistics does not thoroughly deal with all the parts of the data analysis workflow that we have outlined. This is not a criticism of statistics, but just a recognition of its particular focus.

This general point about real world data analysis being more than just the traditional focus of mathematical statistics was actually made decades ago by Tukey (1962). There, Tukey, who was one of the most influential statisticians of the 20th century and a pioneer of exploratory data analysis and data visualization, preferred the term *data analysis* as the general term for what he and other statistical analysts actually do in practice. For Tukey, inferential statistics and statistical modelling was necessary and vital, but only as a component of a much bigger and multifaceted undertaking, which he called data analysis.

While the general spirit of the argument about the breadth and scope of data analysis that Tukey (1962) outlined is very much in keeping with perspective we follow here, modern data analysis has a character that goes Tukey's vision, however broad and comprehensive it was. This is due to the computing revolution. For example, when Tukey was writing in the early 1960s, the world's fastest computers[1] were capable of around 1 million calculations per second. Approximately 60 years later in 2020, the world's fastest computer[2] is capable of around 500 quadrillion (5×10^{17}) calculations per second, and a typical consumer desktop can perform 100s of billions of calculations per second. This revolution has transformed all aspects of data analysis, and now computing is as vital and integral part of data analysis and is mathematics and statistics. It was largely the recognition of the vital and transformative role of computing that lead Cleveland (2001) to coin the term data science. As we use the term, therefore, data science is the blend of computational and statistical methods applied to all the aspects of data analysis.

Even if we accept that the defining feature of data science is generally the combined application of computational tools and statistical methods to data analysis, the term data science has some popular connotations that are somewhat at odds with the more general understanding of the term that we are following in this book. In particular, for some people, data science is all about concepts like big data, machine learning, deep learning, data mining of massive unstructured data sets including natural language corpora, predictive analytics, and so on. It is seen as largely a branch of computer science and engineering, and is something that is done in the Big Tech companies like Google, Amazon, Facebook, Apple, Netflix, Twitter, and so on. It is absolutely true that data science, particularly as it is practiced in industry, heavily avails of tools like machine learning, big data analysis software, and is applied to the analysis of massive unstructured data sets. As real and important as of these activities are, we see them here as just one application of data science as we generally understand the term. Also, in this particular application of data science, some topics and issues take precedence or dominate others. For example, in these contexts, the software and hardware problems of being able to analyse data that is on such a large scale are the major practical issues. Likewise, for some applications, being able to perform successful predictions using statistical methods is the only goal, and so the assumed statistical models on which these predictions are based are less important or even irrelevant (see Breiman 2001 for well known early discussion of these two different general "cultures" of using statistical methods).

In summary, in this book, we use the term data science as the general term for modern data analysis, which is something that always involves a tight integration of computational and statistical methods and tools. In this, we are hopefully faithfully following the broad and general understanding of what real world data analysis entails as described by Tukey (1962), albeit with the additional vital feature of intensive use of computational tools. In some contexts, data science has a more particular focus on big data, data mining, machine learning, and related concepts. That particular focus is not the focus of this book, and so this book may is probably not ideal for anyone keen to learn more about data science in this sense of the term.

[1] https://en.wikipedia.org/wiki/Atlas_(computer)
[2] https://en.wikipedia.org/wiki/Fugaku_(supercomputer)

Why R, not Python?

We have stated repeatedly that computational methods and tools are vital for doing data science. In this book, the computing language and environment that we use is *The R Project for Statistical Computing*, simply known as R. More specifically, we use the modern incarnation of R that is based on the so-called *tidyverse*. In Chapter 2, we provide a proper introduction to R. Here, we wish to just outline why R is our choice of language and environment, what the alternatives are, and what this entails in terms of the our conception of what data science is and how it is practiced.

Given our conception of the data science workflow that we outlined in Figure 1, R is an inevitable choice. We believe that R is simply the best option to perform all the major components that we outline there. For example, for the data wrangling component, which can be extremely labourious, R packages that are part of the tidyverse such as `readr`, `dplyr`, `tidyr` and so on, which we cover in Chapter 3, makes data wrangling fast and efficient and even pleasurable. For data visualization, the `ggplot2` package provides us with essentially a high level and expressive language for data visualization. For the statistical modelling loop, which we cover in all the chapters of Part II of this book, R provides a huge treasure trove of packages for virtually every conceivable type of statistical methods and models. Also, R is the dominant environment, using packages like `rstan` and `brms`, for doing Bayesian probabilistic modelling using the Stan probabilistic programming language. We cover Bayesian models all throughout the chapters of Part II. For communication, R provides us with the ability to produce reproducible data analysis reports using RMarkdown, `knitr`, and other tools, which we describe in Chapter 7.

Everything we cover in this book could be done using another programming language, or possibly using some set of different languages. Chief amongst these alternative is Python. Python is close to being the most widely used general purpose programming language of any kind. It has been very popular for almost two decades, and its dominance and popularity has been increasing in recent years. Moreover, one of Python's major domains of application is data science, with some arguing that it should preferred over R for data science generally. For the big-data, big-tech, data-mining, machine learning sense of data science that we mentioned above, Python certainly ought to be the dominant choice over R. This is for multiple reasons. First, Python is now the principal computing language for doing machine learning, deep learning, and related activities. Also, because Python is a general purpose programming language and one that is widely used on the back end of web applications, this makes integrating data science tools with the "production" web server software much easier and scalable. Likewise, Python is a very powerful and well designed general purpose programming language, which entails that it is easier to write complex highly structured software applications in Python than in a specialized language like R. This again facilitates the integration of Python data science tools with production or enterprise level software applications. Nonetheless, for the more general conception of data science that we are following in this book, Python is more limited than R. For example, for data wrangling of typical rectangular data structures, Python's `pandas` packages, as excellent as it is, is not as high level and expressive as R's tidyverse based packages like `dplyr`, `tidyr`, etc. The entails that wrangling data into shape in R can be easier and involve less lower level procedural and imperative code than when using Python. Likewise for data visualization. Python's `matplotlib` package is very powerful but is also lower level than `ggplot2`. This entails that relatively complex visualization requires considerably more procedural or imperative code, which is harder and slower to read and write than using the more expressive high level code of `ggplot2`. The higher level counterpart of `matplotlib` is `seaborn`, which is excellent, but `seaborn` is less powerful and extensive in terms of its features than `ggplot`. For statistical modelling, at the moment, there frankly is no competition between R and Python. Even though Python has excellent statistics packages like `statsmodels`, these provide only a fraction of the statistical models and methods that are available from R packages. Finally, although dynamic notebooks like Jupyter[3] are widely used by Python users, and are excellent too, it is not as easy to create reproducible reports, for example for publication in scientific journals, using Jupyter as it is using RMarkdown and `knitr`. In fact, currently the easiest way to write a Python based reproducible manuscript is to use Python *within* R using the `reticulate` package.

[3] https://jupyter.org/

Who is this book for?

As mentioned at the start of this chapter, the prototypical audience at whom this book is aimed are those engaged in data analysis in scientific research, specifically research at or beyond PhD level. In scientific research, statistics obviously plays a vital role, and specifically this is based on using data to build and interpret statistical or probabilistic models of the scientific phenomenon being studied. This book is heavily focused on this particular kind of statistical data analysis. As we've mentioned, in data science as it is practiced in industry and business, often the other "culture" of statistics (see Breiman 2001), namely predictive analytics and algorithms, is the dominant one, and so this book is not ideal for those whose primary data science interests are of that kind.

We've explicitly stated that this book is intended for those doing research in the social sciences, but this also requires some explanation. The explicit targeting of the social sciences is largely just to keep some focus and limits to the sets of examples that are used throughout the book. However, beyond the example data sets that are used, there is little about of this content that is of relevance to only those doing research in social science disciplines. All the content on data wrangling, exploration and visualization, statistical modelling, etc., is hopefully just as relevant to someone doing research in some field of biology, as it is to someone doing research in the social sciences. The nature of the data in terms of its complexity, and the nature of the analysis of this data using complex statistical models, is traditionally very similar in biology and social sciences. In fact, the statistics practiced in all of these these fields has all arisen from a common original source, particularly the early 20th century pioneering work of RA Fisher (for example, Fisher 1925).

We assume that the readers of this book will already be familiar with statistics to an extent. For example, we assume that they've taken undergraduate level courses introducing statistics as it is used and applied in some discipline of science. We will present the statistical methods that we cover from a foundational perspective, and so not assume that readers are already confident and familiar with the fundamental principles of statistical inference and modelling. However, we do assume that they will have already had an introduction to statistics so that terms like the normal distribution, linear regression, confidence intervals, and so on, will be relatively familiar terms, even if they don't have a very precise grasp of their technical meaning. On the other hand, we do not assume any familiarity with any computing methods, nor R in particular. In fact, we assume that many readers will be brand new to R, and few will be already very experienced with R.

The style and structure of this book

Apart from this brief introductory chapter, all the remainder of the book is a blend of expository text, R code, mathematical equations, diagrams and R based plots. It is intended that people will read this book while using R to execute all the code examples and so produce all the results that are presented either as R output or as figures. Of course, if readers prefer to read first and the run the code later, perhaps on a second reading, that is entirely a matter of preference. However, all the code that we present throughout this book is ready to run, and does not require anything other than the R packages that are explicitly mentioned in the code and the all the data sets that are being used, which are all available in the website that accompanies this book.

The book is divided into three parts. Part I is all about the parts of the data science workflow shown in Figure 1 except for the statistical modelling loop part. Thus, in Part I we provide a comprehensive general introduction to R, a chapter on data wrangling using `dplyr`, `tidyr` etc., a chapter on data visualization, and another on data exploration. In another chapter, we go into more detail about programming in R, and provide a chapter on doing reproducible data analysis using tools like RMarkdown and Git. Part II of the book, which is the largest part, is all about the statistical modelling loop part of the data science workflow. There, we provide a general introduction to statistical inference, and then cover all the major types of regression models, specifically linear normal regression, generalized linear models, multilevel models, nonlinear regression, and path analysis and related models. In Part III, which is the shortest part, we cover some specialized topics that we not necessarily part of the statistical modelling topics, but not general or introductory either. Specifically, in Part III, we provide an introduction to using R for high performance computing, making interactive graphics web apps using Shiny, and a general introduction to Bayesian probabilistic programming using Stan.

Reference

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3): 199–231.

Cleveland, William S. 2001. "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." *International Statistical Review* 69 (1): 21–26.

Fisher, Ronald A. 1925. *Statistical Methods for Research Workers.* Edinburgh: Oliver; Boyd.

Tukey, John W. 1962. "The Future of Data Analysis." *Annals of Mathematical Statistics* 33 (1): 1–67.