Journal of Statistical Software

October 2017, Volume 81, Book Review 1. doi: 10.18637/jss.v081.b01

Reviewer: Håkon Otneim
Norwegian School of Economics

Business Analytics Using R – A Practical Approach

Umesh R. Hodeghatta, Umesha Nayak
Apress, New York, 2017.
ISBN 978-1-4842-2513-4. 282 pp. USD 39.99 (P), USD 29.99 (e).
http://www.apress.com/9781484225134

Introduction

The plethora of textbooks on business analytics and similar topics has received yet another member, this time by authors Umesh R. Hodeghatta and Umesha Nayak, who employ the subtitles "using R" and "a practical approach" to distinguish their book from the masses. The authors are experienced professionals within the fields of computing, data and machine learning, so I was really looking forward to reading this book. Apart from the usual qualities we demand from a textbook, such as clear language, scientific precision and good examples, I was also hoping that the book would provide interesting perspectives to the *modern* and *practical* use of statistical methods in business applications that could open the eyes of the aspiring student, the industrious consultant and the far-sighted manager to the fantastic potential that lies in understanding and exploiting data. Also, on a more personal level, teaching business statistics for a living, I am always on the lookout for powerful and illuminating arguments to present to my students.

Although I understand, and appreciate, the motivation behind a textbook like this, it is my opinion that the execution of this book fails. The prose is immature, the composition is confusing, the examples are fundamentally uninteresting, the factual content is dubious and at times plainly wrong, the editing is sloppy, and the underlying pedagogical strategy is far beyond my understanding. I realize that these assertions are strong, but I will carefully present my case in the following section.

A review of the contents

The opening chapter "Overview of Business Analytics" is innocent enough. As one would expect from a book like this, the authors try to motivate the value of business analytics through a series of scenarios, such as "You visit a hotel in Switzerland and are welcomed with your favorite drink and dish; how delighted you are!", or "You enter a grocery store and you find that your regular monthly purchases are already selected and set aside for you [...]. How happy you are!", and so on. Definitely awkward, but an honest attempt to illustrate the value of data analysis in the modern world. Well, except for one vital element: there is no *data* in this story! Indeed, section headers like "Survival and Growth in the Highly Competitive World", "Marketing and Sales" and "Product Design" suggest that we are in for a thorough introduction to a rich world of analysis problems that *real* people, in the *real* world, need to tackle. Instead, the authors serve vague and general statements, truisms really, such as "Technology has grown by leaps and bounds over the last few years" and "After-sales service and customer service is an important aspect that no business can ignore". Such perspectives are fine of course, and I totally understand the message here: Modern businesses generate massive amounts of data that, if properly exploited, represent gold mines of information. The chapter is quite dull and vaguely written, though. Instead of motivation, the introduction rather gives a grim foreboding of the stunning lack of substance that the reader has to endure for almost 300 pages. The only pieces of concrete statements to be found concern the alleged qualities of this book, for example that it "[...] takes you through the exciting field of business analytics and enables you to step into this field as well", and that it contains "[p]ractical cases and examples that enable you to apply what you learn from this book". Nothing could be farther from the truth.

Two chapters on R follow next, but the authors have spent very little effort in creating an inspiring introduction to R. After a detailed step-by-step README-style instruction to installing R and **RStudio** in Windows (Mac and Linux user are simply instructed to read the installation instructions for their OS), we enter into a series of toy examples on vector creation and object types. Does a novice in statistics and business analytics really need to store complex numbers, and to construct and subset multidimensional arrays from day one, if ever? Sure, the difference between a matrix and a data frame is important to understand, and factors and character vectors are fundamentally different to an R user. Learning the usage of, and distinction between, `apply()`, `lapply()`, `sapply()` and `tapply()` is absolutely handy in efficient code writing, but in what way are these particular topics helpful to the reader of this text? How can the following code chunk explain the concept and power of the function? (taken verbatim from p. 55):

```
>
> myFunc <- function(){
+ print("My first function")
+ }
> myFunc()
[1] "My first function"
>
```

Indeed, the authors have completely forgotten the title of their book when writing these chapters. The material has nothing to do with business analytics, nor is it a practical approach. It is merely a general primer on the basic programming structures in R that, I suppose, could be useful to some absolute beginners if heavily revised and shortened to a page or two. A much stronger strategy in my opinion would be to *focus on the data*. After all, the reader of this book is probably not very interested in R itself and its intricacies, but rather in how to use R to solve his or her analytical problems. Why not introduce a substantial, real data set along with some non-trivial questions to which we seek answers? What constitutes a tidy data set? How is it loaded into R, and what commands can we use to calculate useful summary statistics, make pretty plots and perform simple analyses like t-tests and regressions, in a way that *helps our understanding of the data*? On such perspectives we are sadly left in the dark by this book.

The ability to produce compelling descriptive statistics is a crucial skill for anyone working with data, especially in business, but the authors do not even try to convey this message in their Chapter 4 on the topic. Instead, they draw up a strange imaginary world in the beginning of the chapter, in which you are stuck trying to cross a river. Alongside the riverbank stands a magical sign that somehow reveals more and more information about the depth of the river, such as the mean, the median and some quantiles. The morality of the story is that, as you have received the information that the 95th percentile is 5 feet,

> "You may now believe that the maximum points may be rare and you can by having faith in God can cross the river safely." [sic]

This is absurd. Besides registering that the headlines are more or less what one would expect in a chapter like this (such as "Mean", "Median", "Quantiles", "Standard deviation", "Histogram", "Scatterplot"), my general impression is that the text is poorly written, and contains absolutely no reference to interesting data (small collections of numbers do not become interesting just because they have dollar signs in front of them, or supposedly represent generic attributes such as "age of workers" or "salary"). Also, on p. 72, the authors, somewhat clumsily, propagate the *spectacular* misconception that "a data set which consists of large number of items is generally said to have a normal distribution or a normal curve", and conclude the chapter with a misplaced section on "Probability", which contains the following pieces of information: "Mutually non-exclusive events are the ones that are not mutually exclusive" (correct, but hardly interesting), and "the probability of n mutually non-exclusive events is $P(A1 \text{ or } A2 \text{ or } \ldots \text{ or } AN) = P(A1 \cup A2 \cup \ldots \cup AN) = P(A1) + P(A2) + \ldots + P(AN) - P(A1 \cap A2 \cap A3 \cap \ldots \cap AN)$,", which is not correct, and far from it.

Chapter 5 on "Business Analytics Process and Data Exploration" starts with a listing of the so-called business analytics life cycle, from Phase 1 (Understanding the business problem) to Phase 8 (Deploy the model), where each phase is discussed quickly, vaguely and in stereotypical terms. For example, Phase 4 (Explore and visualize the data) contain little more than this:

> "This exploration enables you to get a better understanding of the data. You can become familiar with the types of data you collected, identifying outlets and various data types, and discovering your first insights into the data."

Yes, this is true I suppose, but then what? How is this *useful*? What do I *do*? Then, under Phase 6 (Evaluate the model):

> "It is important to evaluate the model, and to be certain that the model achieves the business objectives specified by business leaders. This requires in-depth knowledge of statistics, [...]. Generally, an advanced education such as a PhD may be useful."

I do not understand what the authors seek to achieve by this remark.

The authors decide to display a low-resolution 14 × 14 matrix of scatterplots on p. 113, breaking just about every good practice of descriptive statistics, and they provide a completely incomprehensible section on "Sampling" that out of nowhere concludes that

> "here is a simple formula for calculating a sample: If the population standard deviation is known, then
>
> $$\text{n=(z} \times \text{sigma/E)}\widehat{\ }\,^{2}$$
>
> If standard deviation is unknown
>
> $$\text{n=(p)(1-p)*(z/E)}\widehat{\ }\,^{2}\text{”,}$$

where I have tried, to the best of my abilities, to emulate the appearance of the formulas as they are presented in the book. None of the symbols are ever defined, and the formulas themselves are not justified in any way. It is truly a mystery to me what this is about.

The next five chapters set out to cover some statistical methods: two chapters on supervised and unsupervised learning (which in this book means classification and clustering), and then one chapter each on simple, multiple and logistic regression. This looks fine on the face of it, although perhaps a bit over-simplified and old-fashioned. The contents, on the other hand, makes this book fall utterly apart. Let me give some examples.

Besides giving a superficial description of the classification problem, Chapter 6 contain lengthy step-wise calculations of class probabilities, followed by incomplete low-resolution screenshots on how to do naïve Bayes classification in R using some external packages. On several occasions throughout the book the authors refer to colors in black-and-white figures. For example, in Chapter 6, they "illustrate" the concept of impurity with a figure (p. 139) in which the states of "very impure", "less impure" and "minimum impure" are shown as gray dots in a circle, but the three states look identical without colors. Chapter 7 on classification is mostly data-free, and the explanation of k-nearest-neighbors borders on the comical with a drawing of a rubber duck together with some other types of birds that, I suppose, is meant to illustrate "similarity" or "closeness" between objects. The first step in the algorithm is especially illuminating:

> "1. Select k. It can be 1 or 2 or 3 or anything." (p. 170).

The chapters on regression are full of misconceptions and errors. I recognize *some* important points, such as the motivation behind doing regression, and the role of the correlation coefficient within the regression framework as a measure of linear dependence, but in my opinion they are explained in ways that work *against* the reader instead of helping him/her through the concepts. The following definition of the correlation, found on p. 190, is an example of what I mean:

- Average of [(independent variable in standard units) x (dependent variable in standard units)] (i.e., mean[Σ(z-score of x)*(z-score of y)]).

The examples in these chapters are completely unsuitable to illustrate regression. The response variable is almost a perfect linear function of the explanatory variable(s), leading to R^2-values in the 99.5%-range, which in turn "shows the excellent relationship between the response variable [...] and the independent variable [...]" (p. 196). The implication that R^2

somehow measures the excellence of the regression model (whatever that means) is *fundamentally* wrong and, if consumed with an open mind, quite damaging to the education of the poor reader. It should be noted that other elements of model validation (checking for normality, autocorrelation and so forth) are generally confusing, superficial, vague and misleading.

The treatment of regressions ends in a memorable display of absurdness, as the authors discuss the possibility of multicolinearity in their final example:

> "Typically, a rule of thumb for the multicolinearity test to pass is that the VIF value should be greater than 5. The following test shows the calculation of VIF:
>
> ```
> > vif(attri_logit_model_2)
> Work_Challenging       Work_Envir    Compensation       Tech_Exper
>         1.984868         2.461992        1.611694         1.562850
> ```
>
> As you can see, our model does not suffer from multicolinearity."

The final chapter in the book, "Big data analysis – Introduction and future trends" can be safely ignored. The section headers are simply a long listing of buzzwords (such as "Internet of Things", "Artificial Intelligence", "Vertical and Horizontal Applications", "Cloud, Cloud, Everywhere the Cloud") that shows up in the table of contents on the Amazon listing, but they are followed only by a few general statements without substance.

Some final remarks

I have read every word in this book, and I have considered every part of it seriously. It has lead me to write many bad things, because this is in my opinion a poor quality book. I do not want to put all the blame on the authors, however, because it is the job of the editor to turn the raw manuscript into a final product, and in some cases even turn down manuscripts that do not make the cut (which I believe should be the case here). This book, on the other hand, appears to be completely unedited. All my technical objections aside, I have to say that it is a visual disaster. Unreadable low-contrast screenshots of icons have been inserted in-text, apparently to help the beginner navigate in Microsoft Windows. Many illustrations seem hastily made, and explain very little. There are almost as many different code listing formats as there are code listings. For example, on pp. 26–27, there are three (!) completely different formats for showing code. Instead of being a lubricant that reduces friction when absorbing the content, the editing (or lack thereof) works more like pouring sand into the mental machinery of the reader. I can only conclude that the purpose of this book is not to make beginners "understand and learn the fundamentals of analytics using R" as it says on the cover, but rather to quickly monetize the demand for books on the topic.

If you want an easy-to-read introduction to the exciting world of data analysis, business applications or big data, pick one up while waiting for your next flight; there are plenty well-written paperbacks out there. If you want to learn something new and useful about these topics, please, read another book.

Reviewer:

Håkon Otneim
Norwegian School of Economics
Department of Business and Management Science
Helleveien 30, 5045 Bergen, Norway
E-mail: hakon.otneim@nhh.no

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
October 2017, Volume 81, Book Review 1 *Published:* 2017-10-27
doi:10.18637/jss.v081.b01