An Introduction to Statistical Learning: with Applications in R

Gareth James, Daniela Witten, Trevor Hastie, & Robert Tibshirani

2020-05-20

Contents

Chapter 1

Preface

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation $a^2 + b^2 = c^2$.

The **bookdown** package can be installed from CRAN or Github:

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Remember 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. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.org/tinytex/.

I (not an author) am compiling this book for myself in order to learn both ISLR's material and how to use the **bookdown** package.

Chapter 2

Introduction

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.

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2020) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

Chapter 3

Linear Regression

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 **??**.

This chapter is about *linear regression*, a very simple approach for supervised learning. In particular, linear regression is a useful tool for predicting a quantitative response. Linear regression has been around for a long time and is the topic of innumerable textbooks. Though it may seem somewhat dull compared to some of the more modern statistical learning approaches described in later chapters of this book, linear regression is still a useful and widely used statistical learning method. Moreover it serves as a good jumping-off point for newer approaches: as we will see in later chapters, many fancy statistical learning approaches can be seen as generalizations or extensions of linear regression. Consequently, the importance of having a good understanding of linear regression before studying more complex learning methods cannot be overstated. In this chapter, we review some of the key ideas of the linear regression model, as well as the least squares approach that is most commonly used to fit this model.

Recall the **Advertising** data from Chapter 2. Figure 2.1 displays **sales** (in thousands of units) for a particular product as a function of advertising budgets (in thousands of dollars) for **TV**, **radio**, and **newspaper** media. Suppose in our role as statistical consultants we are asked to suggest, on the basis of this data, a marketing plan for next year that will result in high product sales. What information would be useful in order to provide such a recommendation? Here are a few important questions that we might seek to address:

1. *Is there a relationship between advertising budget and sales?*
   Our first goal should be to determine whether the data provide evidence of an association between advertising expenditure and sales. If the evidence is weak, then one might argue no money should be spent on advertising!

2. *How strong is the relationship between advertising budget and sales?*
   Assuming there is a relationship between advertising and sales, we would like to know the strength of that relationship. In other words, given a certain advertising budget, can we predict sales with a high level of accuracy? This would be a strong relationship. Or is a prediction of sales based on advertising expenditure only slightly better than a random guess? This would be a weak relationship.

3. *Which media contribute to sales?*
   Do all three media - TV, radio, and newspaper - contribute to sales, or do only one or two of the media contribute? To answer this question, we must find a way to separate out the individual effects of each medium when we have spent the mony on all three media.

4. *How accurately can we estimate the effect of each medium on sales?*
   For every dollar spent on advertising in a particular medium, by what amount will sales increase? How accurately can we predict this amount of increase?

5. *How accurately can we predict future sales?*
   For every given level of television, radio, and newspaper sales, what is our prediction for sales, and what is the accuracy of this prediction?

6. *Is the relationship linear?*
   If there is an approximately straight-line relationship between advertising expenditure in the various media and sales, then linear regression is an appropriate tool. If not, then it may still be possible to transform the predictor of the response so that linear regression can be used.

7. *Is there synergy among the advertising media?*
   Perhaps spending $50,000 on television advertising and $50,000 on radio advertising results in more sales than allocating $100,000 in either television or radio individually. In marketing, this is known as a *synergy* effect, while in statistics it is called an *interaction* effect.

It turns out that linear regression can be used to answer each of these questions. We will first discuss all of these questions in a general context, and then return to them in this specific context in Section 3.4.

3.1 Simple Linear Regression

Simple linear regression lives up to its name: it is a very straightforward approach for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is an approximately linear relationship between X and Y. Mathematically we can write this linear relationship as

$$Y \approx \beta_0 + \beta_1 X \tag{3.1}$$

You might read "≈" as "*is approximately modeled as*". We will sometimes describe (3.1) by saying we are *regressing* Y on X (or Y *onto* X). For example, X may represent **TV** advertising and Y may represent **sales**. Then we can regress **sales** onto **TV** by fitting the model:

$$\textbf{sales} = \beta_0 + \beta_1 \times \textbf{TV}$$

In Equation (3.1), β_0 and β_1 are two unknown constants that represent the *intercept* and the *slope* terms in the linear model. Together, β_0 and β_1 are known as the model *coefficients* or *parameters*. Once we have used our training data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for our model parameters, we can predict future sales on the basis of a particular value of TV advertising by computing

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3.2}$$

where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. Here we use the *hat* symbol, $\widehat{\ }$, to denote the estimated value of an unknown parameter or coefficient, or to denote the predicted value of the response.

3.1.1 Estimating the Coefficients

Bibliography

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

Xie, Y. (2020). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.18.