Multiple Regression Analysis

Bret Hart

October 14, 2016

Abstract

The aim of this report is to reproduce some of the graphical, statistical, and tabular results displayed in chapter 3.2, *Multiple Linear Regression*, of **An Introduction to Statistical Learning**. Referred to as **ISLR**, the textbook is a manifesto of Machine Learning and Linear Models, teaching the material in an approachable yet sophisticated way. In addition, the data used to generate all of the graphs, plots, etc. in the text are freely available - advancing and standing for the tenants of reproducible research, even in a textbook. We seek to create an automated repository which can recreate the findings that they display, using the same data set.

Introduction

The data set which we are studying is an Advertising data set - it is a collection of money spent in 200 different markets on Advertising, with each market's corresponding Sales figures. The data includes information on Television, Radio, and Newspaper advertisement, and for this multiple linear regression project, we will be considering how all three play a role in determining and predicting their corresponding Sales figures. We would like to determine whether there is a meaningful, significant relationship between the three advertisement mediums and Sales. Using these results, we would like to be able to predict future Sales figures based on potential amounts of Advertisement expenditure. Ultimately, we would like to make sophisticated, informed decisions on how to form an Advertising plan in the future, with all of the three possibly confounding variables considered simultaneously and in relation to one another. We want to model this relationship effectively and correctly, and use the model to predict future sales and create a profitable Sales plan.

Additionally, we seek to reproduce certain pertinent figures and tables shown in chapter 3.2 of **ISLR**, intermediary statistical explanations which help us to better understand all the ways in which the predictors and dependent variable are related. The tables, thus, are valuable both because making them proves the reproducibility of the **ISLR** text, but also because the tables do truly help us to understand the Advertising data set better.

Data

More specifically, the Advertising data sets contains **Sales** (in thousands of units) of a particular product in 200 different markets, supplemented by advertising budgets (in thousands of dollars) for the products in three different forms of media: **TV**, **Radio**, and **Newspaper**. For this project we are going to focus on the relationship between the three collected predictors, **TV**, **Radio**, **Newspaper**, and their response, **Sales** - for the purposes of specifically reproducing the figures and findings in **ISLR** and to better understand multiple linear regression.

Methodology

As stated previously, we are focusing on the three advertising mediums of **TV**, **Radio**, and **Newspaper** and their relationship with **Sales**. We will consider both their individual, 1-1 relationship with **Sales**, as well as their combined predictive effect. To do this, we will assume and use the multiple linear model:

$$\textbf{Sales} = \beta_0 + \beta_1\textbf{TV} + \beta_2\textbf{Radio} + \beta_3\textbf{Newspaper}$$

To estimate the coefficients β_0, β_1, β_2, and β_3, we fit a regression model via the multiple least squares criterion.