Simple Regression Analysis

Bret Hart

October 7, 2016

Abstract

The aim of this report is to reproduce the main graphical and statistical results displayed in chapter 3.1, *Simple Linear Regression*, of **An Introduction to Statistical Learning**. Referred to as **ISLR**, the textbook is a manifesto to 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 and each market's corresponding Sales figures. While the data also includes information on Newspaper and Radio advertisement, for the purpose of this project, we are going to focus on Television advertisement expenditure. We would like to determine whether there is a meaningful, significant relationship between TV advertisement and Sales, and, using these results, predict future Sales figures based on amounts of Advertisement expenditure. Ultimately, we would like to make sophisticated, informed decisions on how to form an Advertising plan in the future. We want to model this relationship effectively and correctly, and use the model to predict future sales and create a profitable Sales plan.

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, however, we are going to focus primarily on the relationship between **TV** and **Sales**, for the purposes of specifically reproducing the figures and findings in **ISLR**.

Methodology

As stated previously, we are focusing on the advertising medium of **TV** and its relationship with **Sales**. To do this, we will assume and use the simple linear model:

$$\textbf{Sales} = \beta_0 + \beta_1 \textbf{TV}$$

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

Results

We estimate the regression coefficients via the least squares method in this table:

Some statistics of the least squares model are presented in this table:

Lastly, here is the scatterplot of the mapped Television vs. Sales values, with fitted regression line:

Table 1: Information about Regression Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.03	0.46	15.36	0.00
TV	0.05	0.00	17.67	0.00

Table 2: Regression Statistics

	Statistic	Value
1	RSE	3.26
2	RSS	2102.53
3	R2	0.61
4	F-stat	312.14

```
#install.packages("png")
library(png)
```

```
## Warning: package 'png' was built under R version 3.2.5
```

```
require(png)

x.png <- readPNG("../images/scatterplot-tv-sales.png")
png(x.png)
```