# Simple Regression Analysis

Bret Hart

October 31, 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 $\beta_0$ and $\beta_1$ we fit a regression model via the least squares criterion.

# Results

We estimate the regression coefficients via the least squares method in Table 1:

Table 1: Information about Regression Coefficients

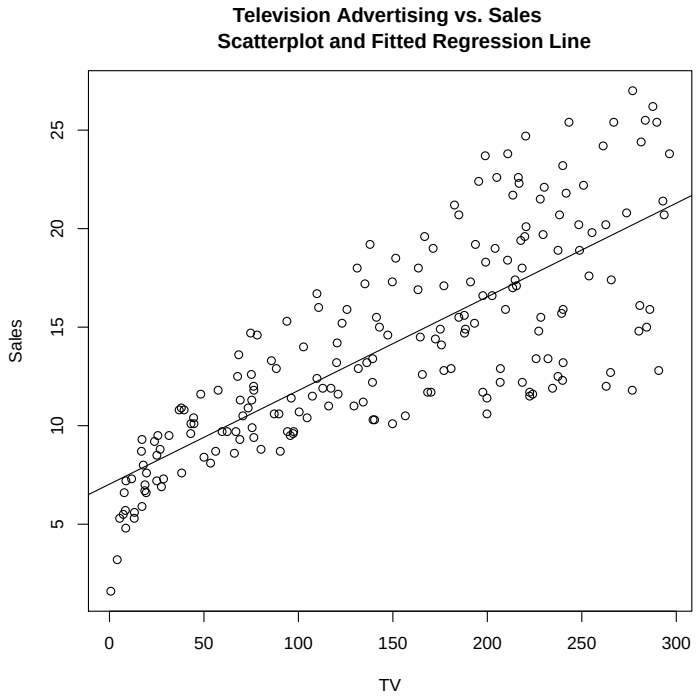
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 7.03 | 0.46 | 15.36 | 0.00 |
| TV | 0.05 | 0.00 | 17.67 | 0.00 |

Some statistics of the least squares model are presented in Table 2:

Table 2: Regression Statistics

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

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




# Conclusions

Based on the extremely high t-value for **TV**, it can be surmised that, when other variables are not accounted for, **TV** advertisement has a significantly large chance of having an effect on **Sales**. Of course, a conclusion based solely on a t-statistic in a vacuum will be innacurate, as the other variables (**Newspaper**, **Radio**) likely influence the markets that **TV** advertising is taking place in to a substantial degree. In addition to the t-statistic, the F-statistic is extremely large as well, but it holds less meaning when there are not multple predictors being matched to a dependent variable. Although its relevancy is questionable, having a significant F-statistic surely helps.

Having a low RSE and realistically substantial R2 value, though, add a good deal of certainty to the reasonability of a statistically relevant fit. While neither statistics are showcasing an extremely close fit, they suggest good things about the goodness of fit, the correlation between **TV** and **Sales**, and substantially bolster the argument that **TV** is an important predictor in determining product sales in a market.

Some interesting observations are that, as TV advertising increases, Sales becomes more variable, in a very observable fashion. Perhaps these markets have subsequently less (or more) advertisement spending in the other mediums, influenced by the increase in TV spending, which leads to these more wild

and unpredictable Sales figures. Additionally, there seem to be a pretty equal number of points above and below the line, but a general trend upwards. This is no detriment to a linear model, as the general upward trend is far more signifcant than spread out values.

Ultimately, increasing **TV** spending seems to have a statistically significant effect on an increase in Sales, but a realistic spending model will take into account how much these product sales really mean in response to a significant but not very positive relationship. The coefficient of **TV** in the regression line is pretty small, so although advertising does seem to have a net positive effect, there is definitely going to be a drop-off in how lucrative increasing expenditures will be past a certain point. Eventually advertising spending will overshadow actual sales revenue, so the acknowledgement of a significant but not very positive relationship can help color marketing decisions in the future. Additionally, it would be wise to start measuring the effects of the other predictors in tandem with **TV**, as these numbers are surely not telling the whole story.