Simple Regression Analysis

Lydia Maher

October 6, 2016

Abstract

A report detailing the performance of a simple regression analysis, based on the examples given in chapter 3 of the book **An Introduction to Statistical Learning**. In this report, we reproduce some of the graphics and statistics included in that chapter.

Introduction

The overarching goal of the project is to increase sales of the product. In order to do this, we undertake analysis to see whether there is a concrete relationship between sales and amount of money spent on advertising. This advertising is split into three different media formats: TV, Radio and Newspaper. By developing a model which predicts sales based on funds allotted to each media domain, we can determine the optimal allocation of funds and achieve the greatest amount of sales.

Data

We are using the dataset given in the book entitled "Advertising.csv". It contains the data for 200 different markets, with the amount spent on TV, Radio and Newspaper advetising (in thousands of dollars) and the amount of sales this spending produced (in thousands of units).

Methodology

In this project, we are focuing on the media of TV and analysing its relationship with Sales. To carry out this analysis, we are using the simple model: Sales = (Beta0) + (Beta2)TV. These coefficients are estimted with regression and least-squares in R.

Results

Here is a table with the resulting co-efficient values:

```
            Table 1:Information about Regression Coefficients
               Estimate  Std. Error  t value    Pr(>|t|)
(Intercept) 7.03259355 0.457842940 15.36028 1.40630e-35
TV          0.04753664 0.002690607 17.66763 1.46739e-42
```

This tells us that our estimate for Beta0 is 7.03 and for Beta1 it is 0.0475. This means that for every additional $1000 spent on TV advertising, we would expect approximately 47.5 additional TVs to be sold (and that below having $7000 spent on advertising, we wouldn't expect to see any effects with zero TVS being expected to be sold under this threshold).

Further analysing our least squares model, it is useful to look at some quality indices:

Table 2: Regression Quality Indices

Quantity	Value
RSS	3.2586564
R2	0.6118751
F-stat	312.1449944

We have a fairly large RSS here (3.26) and this indicates that whatever the true values of the coefficients, any prediction of sales is likely to be off by 3,260 units on average. As the mean value of sales over all markets is approximately 14,000, this gives us a percentage error of 23% (An Introduction to Statistical Learning, pg 69). Our R2 statistic gives us the proportion of variability in Sales that can be explained by TV. This shows that approximately two-thirds of Sales variability is caused by a linear regression on TV. Finally, our F-Stat gives us a highly significant p-value and so we can conclude that the difference in means between Sales and TV is statistically significant.

Here is a scatterplot detailing the showing the data and how it compares to our regression line (error for each point also drawn onto the graph):

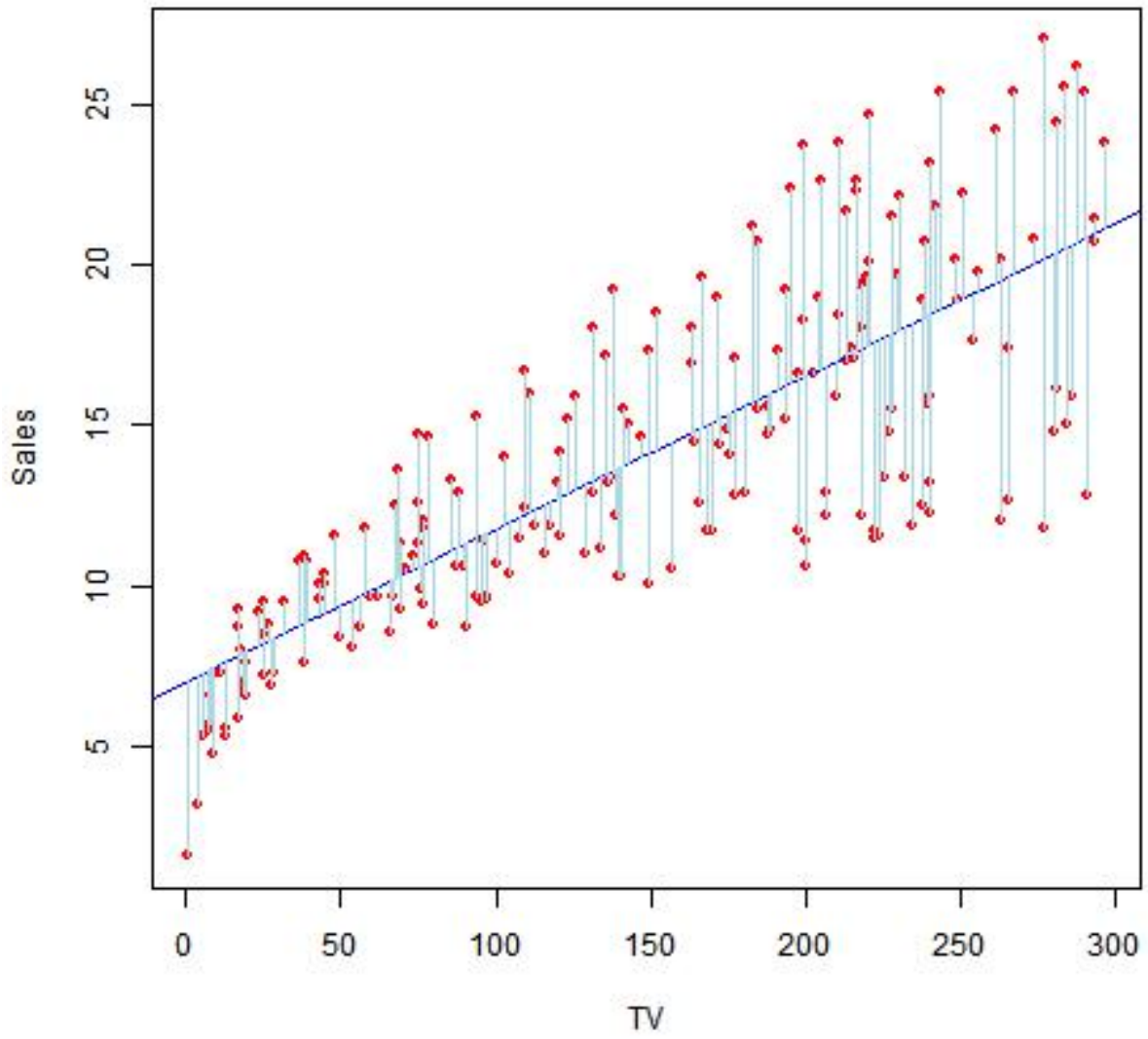

Figure 1: Scatterplot with fitted regression line

This shows graphically that there is quite a lot of error, but also a positive correlation.

Conclusions

According to our data, there is a positive correlation between TV and Sales of about 0.6. However, our model is by no means perfect and the simple regression has a lot of error as shown by the RSS statistic of 3.26. It would be useful if we had more data to be able to refine our model. Based on this simple analysis though, it would seem that more spending on TV advertising will result in more sales.