Abstract

The purpose of the report is to apply computational toolkit to reproduce a simple regression analysis in section 3.1 - Simple Linear Regression of the book "An Introduction to Statistical Learning" by Gareth James, Deniela Witten, Trevor Hastie and Robert Tibshirani. In this report, I used the data set for advertising to perform simple linear regression of Sales on TV budget. I will introduce the steps and tools to work on the reproducible workflow in the following sections.

Introduction

In this report, I will try to find out the relationship between the TV advertisement budget and sales to see if the increase in TV budget will increase the sales of a certain product. I will also do some analysis to see how strong the relationship is between TV budget and sales. By interpreting the key statistics, I can find out the relationship between TV budget and sales and how strong it is.

Data

The advertising data set is downloaded from http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv. The data set includes the advertising budget (in thousands of dollars) for TV, Radio, Newspaper and the Sales (in thousands of units) of a certain product in 200 different markets. In this report, we only need to use the data of TV budget and Sales to find if there's a relationship between them.

Methodology

To analyze the relationship between the Sales and TV budgets, I used the simple linear regression model: $Sales = \beta_0 + \beta_1 * TV$. I used the ordinary least squares estimation to estimate the coefficients and to fit a line for the data set.

Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.0326	0.4578	15.36	0.0000
TV	0.0475	0.0027	17.67	0.0000

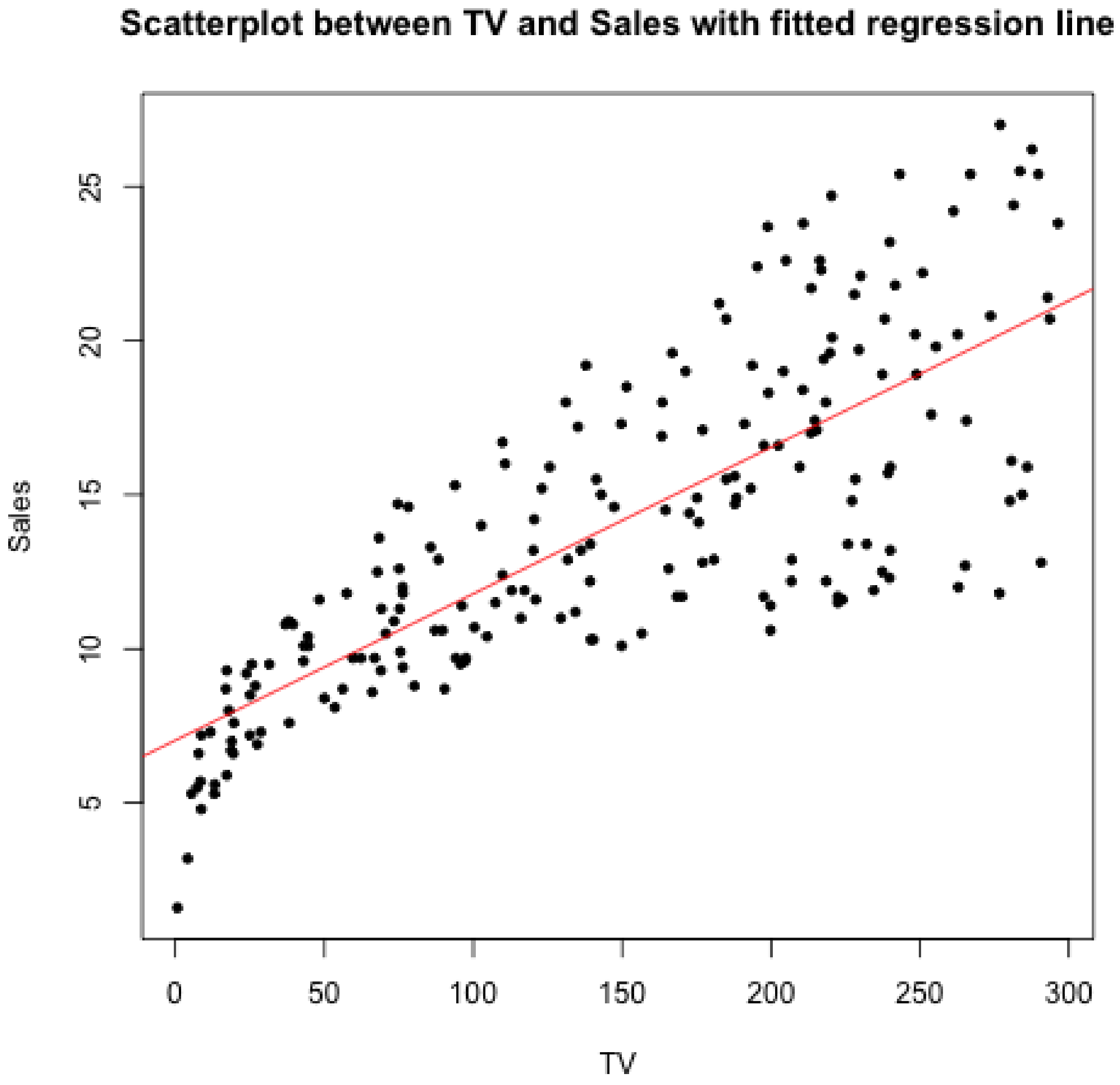
Table 1: Regression Coefficients

From the table above, we can see that the estimated slope is 0.0475. We can interpret the statistics as for every addtion one thousand dollars spent on TV budget, the sales will increase for 47.5 units. From the table, we can also see that the t-values for the estimated slope and estimated intercept are large while the p-values for them are small. Thus, we can conclude that there's a linear relationship between the TV budget and Sales due to the significance of the statistics.

In the table below, we can see that the RSS is 3.26, which is relatively small, which means the simple linear regression model is a relatively good fit. We can also see from the table that R^2 is 0.61, which can be interpreted as 61% of the variation in Sales can be explained by the change in TV budget. Moreover, the large F-statistic further demonstrates that a simple linear regression model is a good fit for the data set and there exists a relatively strong relationship between TV advertising budget and Sales.

	Quantity	Value
1	RSS	3.26
2	R-square	0.61
3	F-Statistics	312.14

Table 2: Regression Quality Statistics



From the graph we can see the positive relationship between TV budget and sales in a more explicit way. However, from the scatterplot we can also see the data is heteroskedastic, which means that we may need to transform the original data or try to fit some other models to find the best fitted model for the data set.

Conclusion

By fitting in the linear regression model and interpreting the statistics, we can conclude that there exists a relatively strong positive linear relationship between the TV budget and Sales. We were also able to replicate the models, graphs and tables in section 3.1 of the book by using tools including Makefile, R, Git, Github, Rmarkdown and so on.