Simple Regression Analysis

Shuotong Wu

10/31/2016

1 Abstract

In this report, we reproduce the main results displayed in section 3.1 Simple Linear Regression (chapter 3) of the book An Introduction to Statistical Learning.

2 Introduction

The overall goal is to provide advice on how to improve sales of the particular product. More specifically, we are assessing the relationship between advertising and sales. If the analysis shows that there is an association between advertising and sales, we will develop a model to predict sales based on advertising budgets.

3 Data

The advertising dataset consists of Sales(in thousands of units) of a particular product in 200 different markets, along with advertising budgets (in thousands of dollars) for the product in each of those markets for three different media (TV, Newspaper and Radio). We will focus on TV budget in this report.

4 Methodology

Since we are studying the relationship between Sales and a single media TV, we will try to use a simple linear model:

$$Sales = \beta_0 + \beta_1 TV$$

We consider Sales and TV advertising budgets in our dataset and try to fit them in a simple linear regression model:

5 Exploration

We compute the regression coefficients in Table 1 below:

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

Table 1: Table 1: Information about Regression Coefficients

As we can see in Table 1, we have a slope of 0.0475, a positive number which suggests that the more budget we put in TV advertising, the more sales we get. Also, we notice that the p values are nearly 0, which means that it would be extremely rare to get a result as unusual as this if the coefficient were really 0.

More information about the least squares model is given in the table below:

	Quantity	Value
1	Residual standard error	3.26
2	R Squared	0.61
3	F-Statistics	312.14

Table 2: Table 2: Regression Quality Indicies

As we can see in Table 2, for our model, residual standard error is 3.26 and R-squared value is 0.61, which is not super high. It indicates that the data is kind of close to the fitted line but our may not be the best model. It suggests that we should try other linear models in the future work.

Below is the scatterplot with fitted regression line. And we can see that as budget for TV advertising increases, the deviation between data and fitted regression line also increases, which also suggests that our model performs worse when TV advertising budget gets large.

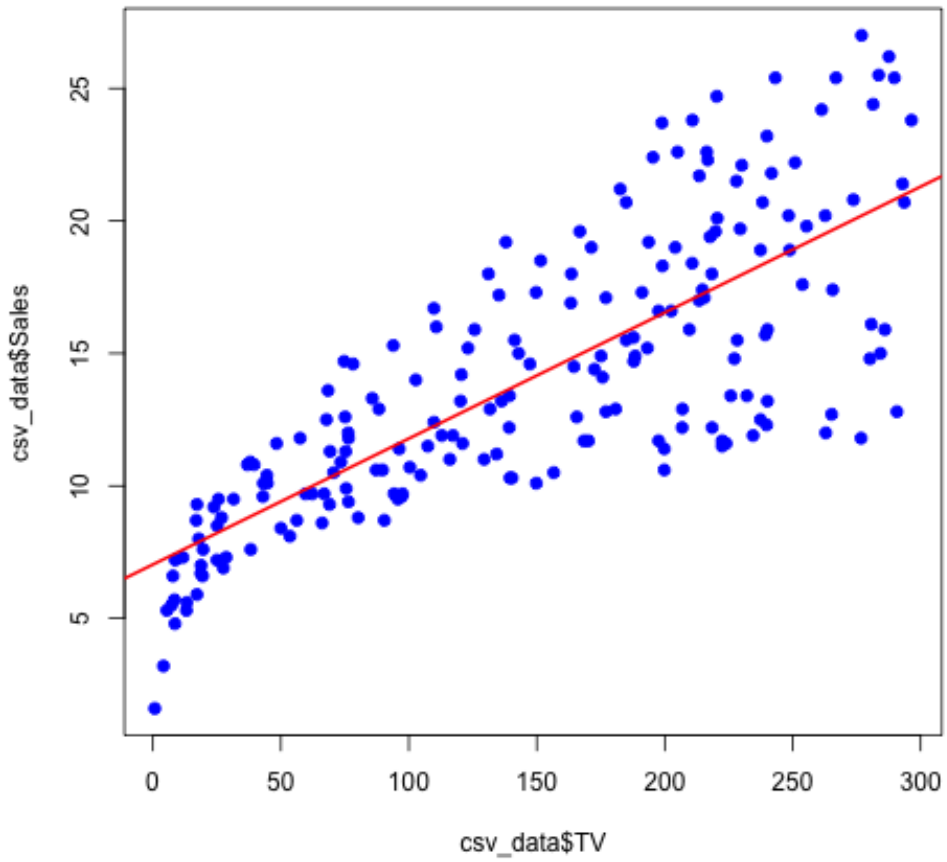

Figure 1: Scatterplot

6 Conclusion

Our simple linear regression model with least square criterion perform OK with our current dataset. However, it may not be reliable when the budget is really high. We should gather more data to train our model. Also we should explore other models for comparison and maybe introduce ensemble method for better quality.