Simple Regression Analysis

Youngshin Kim

Oct 07, 2016

Abstract

In this report, we aim to reproduce some of the results used in the book "An Introduciton to Statistical Learning". To do this, we primarily use R script files for reading/manipulating data and Makefile to automate necessary conversion tasks.

Introduction

From the main data set, Advertising, we use the data on TV marketing and Sales to check if there is any significant relationship between the two variables.
We want to know how Sales changes when the amount of TV marketing changes.
This is not necessarily to check for causality, because correlation doesn't mean causality all the times.

Data

The Advertising data consists of values on Sales (in thousands of unit) of some product in 200 different markets and values on advertising budget (in thousands of dollars) for three different kind of marketing strategies: TV, Radio, Newspaper. For the purpose of this project, we are just interested in the advertising budget for TV.

Methodology

We use simple linear model to model the relationship between TV and Sales. The response variable here is Sales and the predictor variable is TV.
The regression equation is Sales = Intercept + Coefficient * TV. With this equation in mind, we use the lm function in R to find the relevant information for our regression. The syntax for this code is lm(Sales ~ TV, data = advertising) where object 'advertising' is the main data from Advertising.csv. The basic idea behind this function is that we want to find the intercept and the regression coefficients that minize the residual sum of squares.

Results

When we compute the regression coefficients using lm function, we get the following summary statistics

	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

Since the p-values of the intercept and the regression coefficient are close to zero, we have evidence that TV and Sales is positively correlated. The beta coefficient for TV, however, is very small. To be exact, it tells us that when TV advertising budget increase by one unit, we can expect to see an average increase in Sales by 0.05. This is not too large.

	Value
RSS	3.26
R^2	0.61
F-statistic	312.14

Table 2: Values extracted from lm function summary

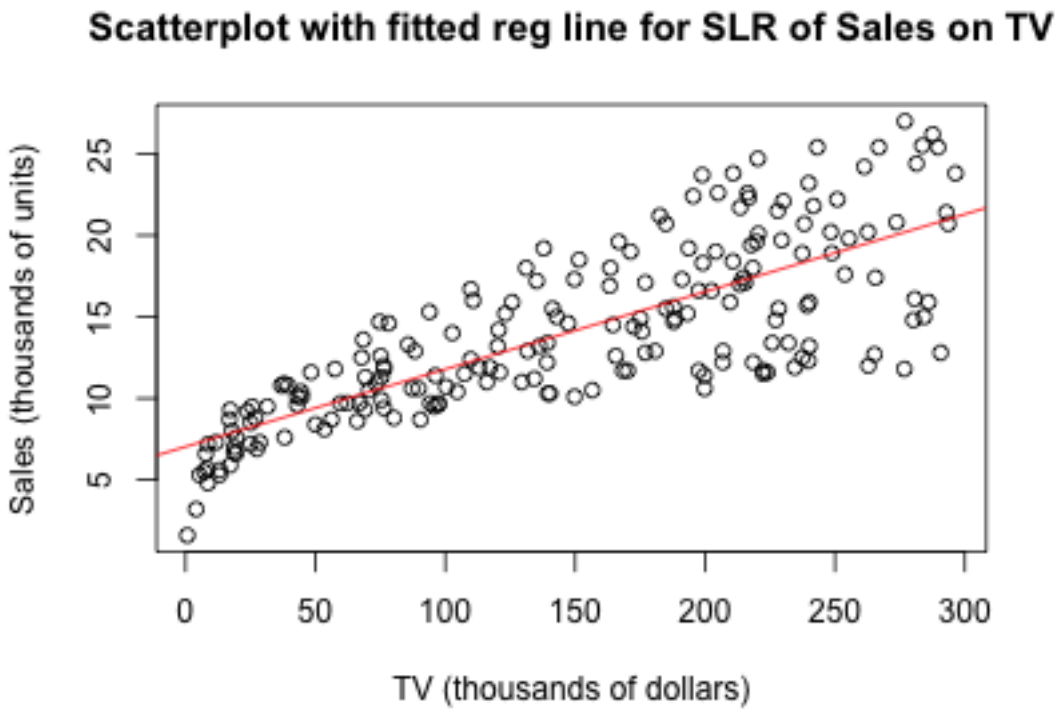

Figure 1: Scatterplot with regression line

This is the scatterplot with regression line fitted to the plot.
It appears that there is some linear relationship between TV and Sales, because the red regression line seems to express the general direction of the relationship between TV and Sales.

Conclusion

The Advertising data had three predictors: TV, newspaper, and radio.
Here we only looked at TV. If we perform multiple regression including all three predictors, we could get a different result for the regression coefficient for TV. In this simple linear regression at least, the results we got seem to tell us that when we spend more money on TV advertising, we can expect to see a mild growth in Sales.