

Simple Regression Analysis

Liang Hao

10/7/2016

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*.

Introduction

The overall goal is to provide advice on how to improve sales of the particular product. More specifically, the idea is to determine whether there is an association between advertising and sales, and if so, develop an accurate model that can be used to predict sales on the basis of the three media budgets.

Data

The Advertising data set consists of the *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*, *Radio* and *Newspaper*. In this report we focus on the possible relation between *TV* and *Sales*. Following is the table for summary statistics for both *TV* and *Sales*:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Variance	Standard Deviation
TV	0.70	74.38	149.80	147.00	218.80	296.40	10831.30	104.07
Sales	1.60	10.38	12.90	14.02	17.40	27.00	27.22	5.22

Table 1: Summary Statistics

From the table above, the relation between the two variables are not very clear yet. We then explore with the following methodology.

Methodology

We consider one media from the data set, *TV*, and study its relationship with *Sales*. For this purpose, we use a simple linear model:

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

To estimate the coefficients β_0 and β_1 we fit a regression model via the least squares criterion.

The null hypothesis here is that the *TV* would not have an effect on *Sales*, and the alternative hypothesis is that *TV* does have an effect on *Sales*.

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

Table 2: Information about Regression Coefficients

Results

After fitting the data to a simple linear regression model, we compute the regression coefficients:

From the table above we can extract the intercept and slope for future graphing. And we see that the p values for both intercept and TV returned by this simple linear regression model are both smaller than 0.05, giving us enough evidence for the alternative hypothesis that these two factors do make a difference.

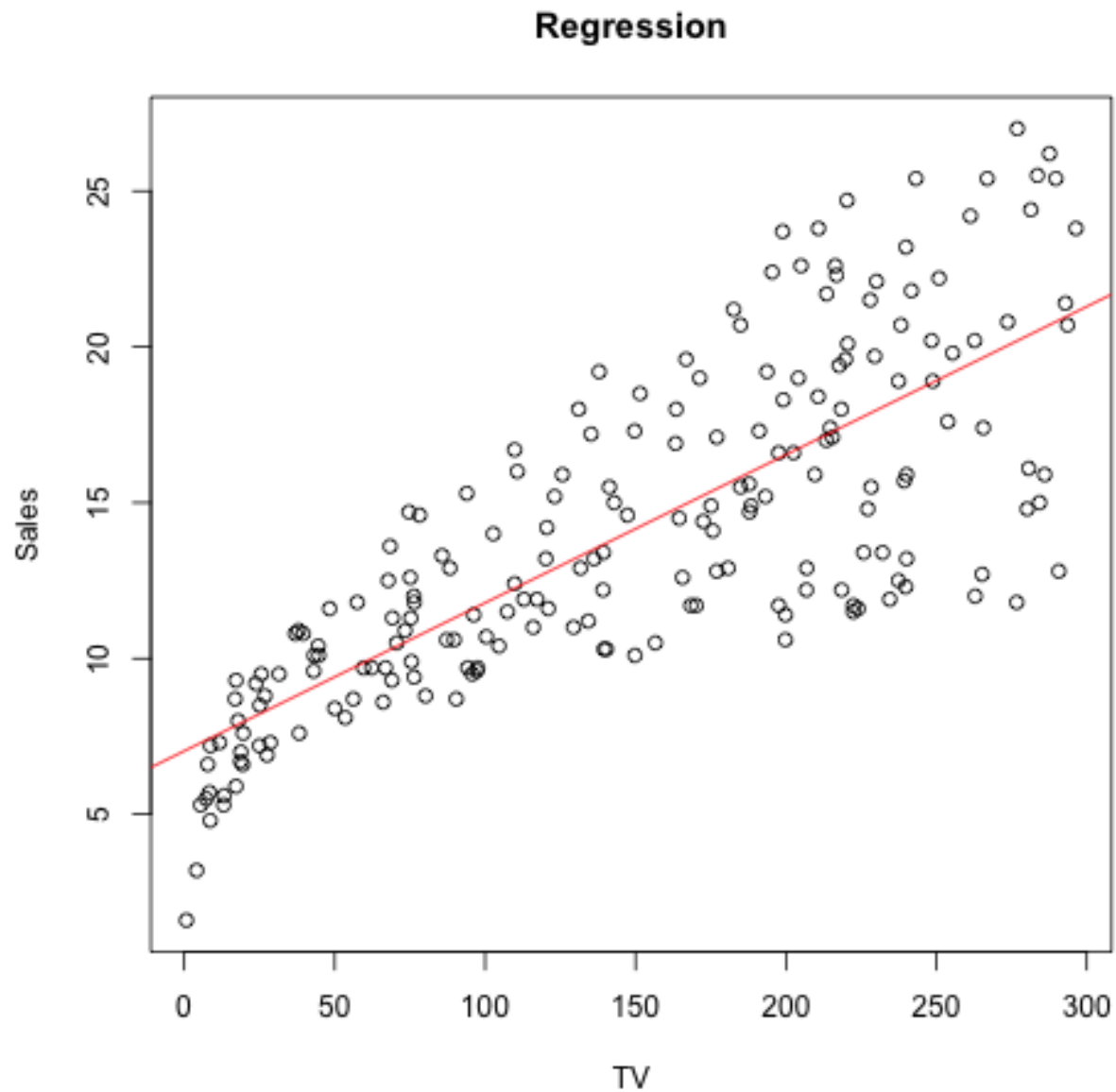
More information about the least squares model is given in the table below:

	Quantity	Value
1	RSS	3.25865636865046
2	R2	0.611875050850071
3	F-stat	312.144994372713

Table 3: Regression Quality Indices

We can see that the Residual Sum of Squares and r squared for this model is relatively small, meaning that the simple linear regression model is a relatively good fit of the data.

And we plot the scatter plot with the fitted regression line.



As we can see from the scatter plot above, the regression line approximates most the relation between TV and $Sales$.

Conclusions

From the analysis above, we may see that TV does have an effect on $Sales$, and the simple linear regression does a relatively good job in capturing such relation.

Thinking ahead, we may consider incorporating other factors in the model to have a even better approximation on how it will go, with the hope that we can make more accurate predictions and thus decisions.