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 of this analysis is to provide advice on how to improve sales of the particular product given the current information. 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. For this analysis specifically, we primarily consider using simple linear regression.

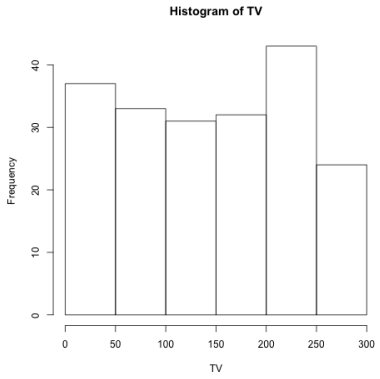
Data

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*:

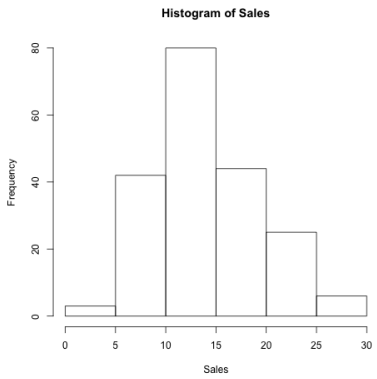
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Variance	Standard Deviation
TV	0.70	74.38	149.80	147.00	218.80	296.40	10831.30	104.07
Sales	1.60	10.38	12.90	14.02	17.40	27.00	27.22	5.22

Table 1: Summary Statistics

From the table above, the relation between the two variables are not very clear yet. We may also have a look at the histograms of their distribution:



From the *Histogram of TV* above, see that the frequencies over the range are approximately even, with *200-250* highest and *250-300* lowest. We then inspect the distribution of *Sales*:



From the *Histogram of Sales* above, we can see an approximately bell shape distribution, with *10-20* the highest, which we could infer from the summary statistics above. We then explore with the following methodology.

Methodology

We consider one media from the data set, TV, and study its relationship with *Sales*. 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*. 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. If the relation did not exist, we would expect β_1 would be close to 0, or the distribution of *Sales* is relatively independent of the distribution of TV.

Results

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

	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

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, showing statistical significance.

Also, the *Standard error* of the two parameters are significantly smaller than the actual values of the parameters.

Therfore, we may have enough evidence against the null hypothesis that the two factors are not related. TV does have an effect on *Sales*.

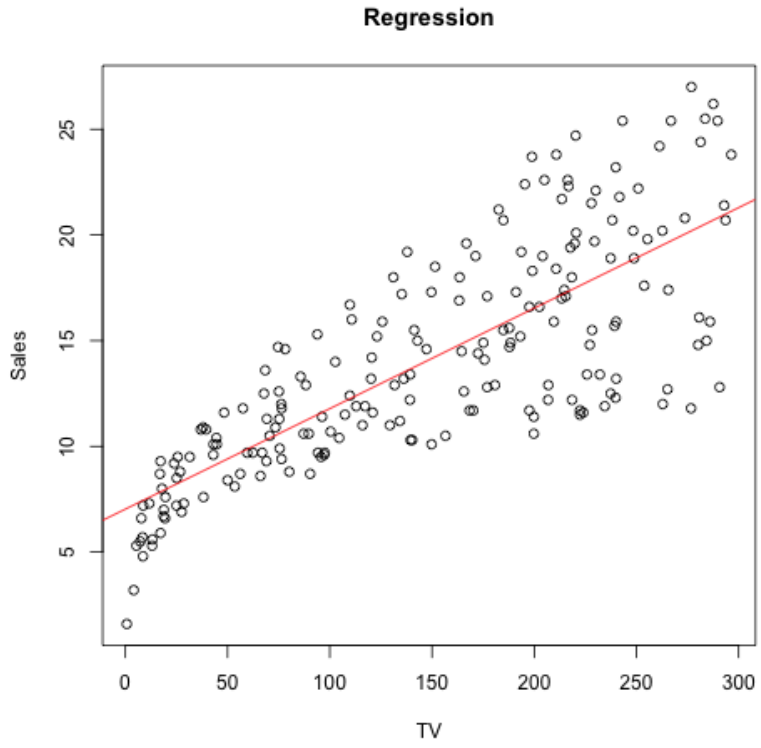
Futhermore, we can dig deeper into the parameters of the least squares model. The table below shows information about a few important indicators when evaluating a model:

	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 scattor plot with the fitted regression line.



As we can see from the scattor plot above, the regression line approxiates most the 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.