Simple Regression Analysis

Siyu Chen

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 explore whether increasing TV advertising budgets will improves sales. In this report, we specifically look at how TV advertising budgets affect sales by calculations and plotting graphs. If a relationship exists between TV advertising budgets and sales, then we want to build a good linear model that can be used for sales prediction based on TV advertising budget.

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). But we will only explore the association between TV advertising budgets and sales in this report.

Methodology

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

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

In order to estimate two coefficients β_0 and β_1, we fit the linear regression model via the least square criterion.

Exploration

First, let's explore our data.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
TV	0.70	74.38	149.80	147.00	218.80	296.40
Sales	1.60	10.38	12.90	14.02	17.40	27.00

Table 1: Data Summary

As we can see, we have several statitics, including Min, 1st Quatile, Median, Mean, 3rd Quatile and Max. We can get some initial impressions about data but we need to explore more in depth.

Then, let's take a look at their histograms.

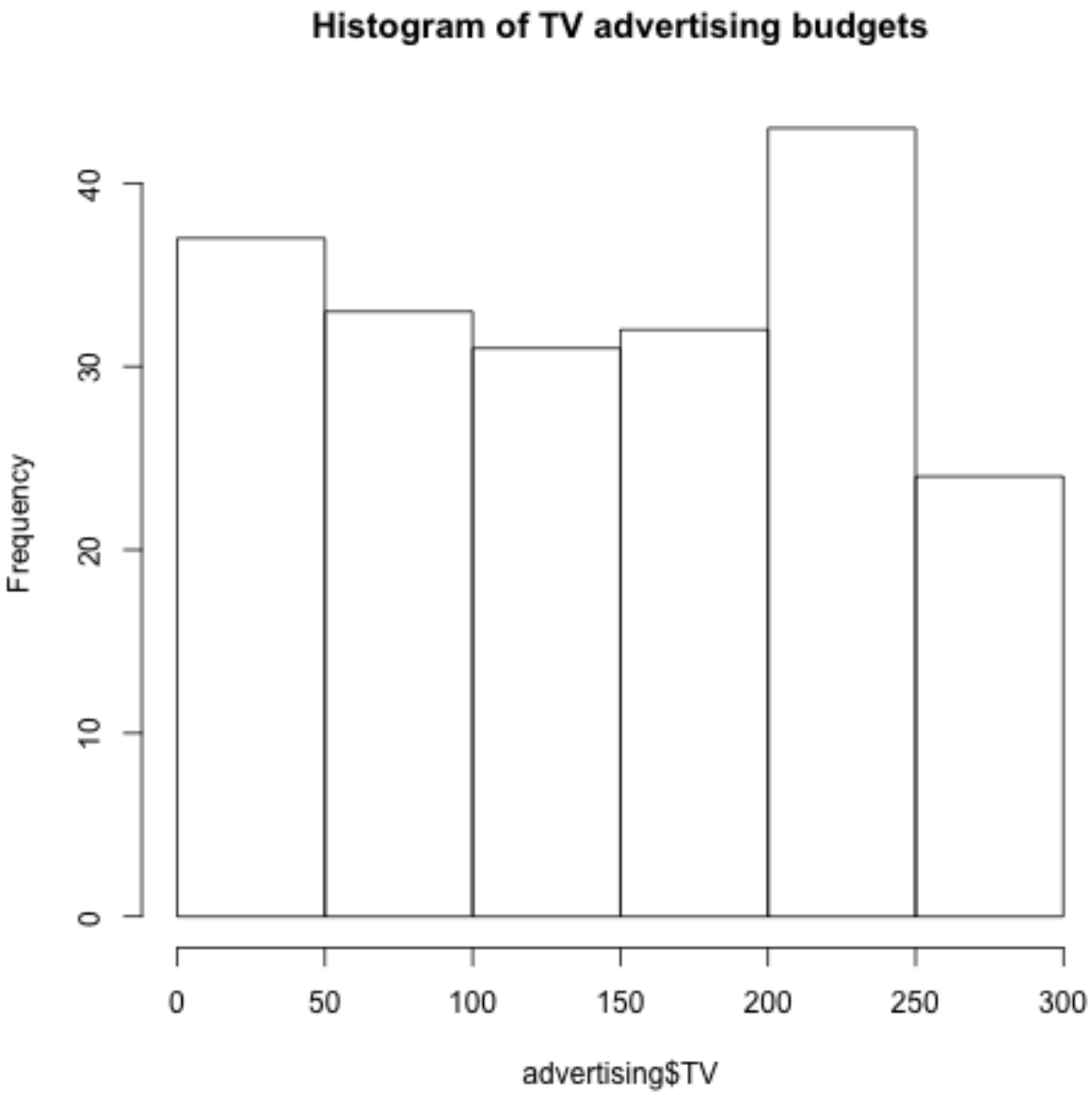

Figure 1: Histogram TV Avertising Budgets

As we can see, the frequencies of TV advertising budgets are approxiaately even.

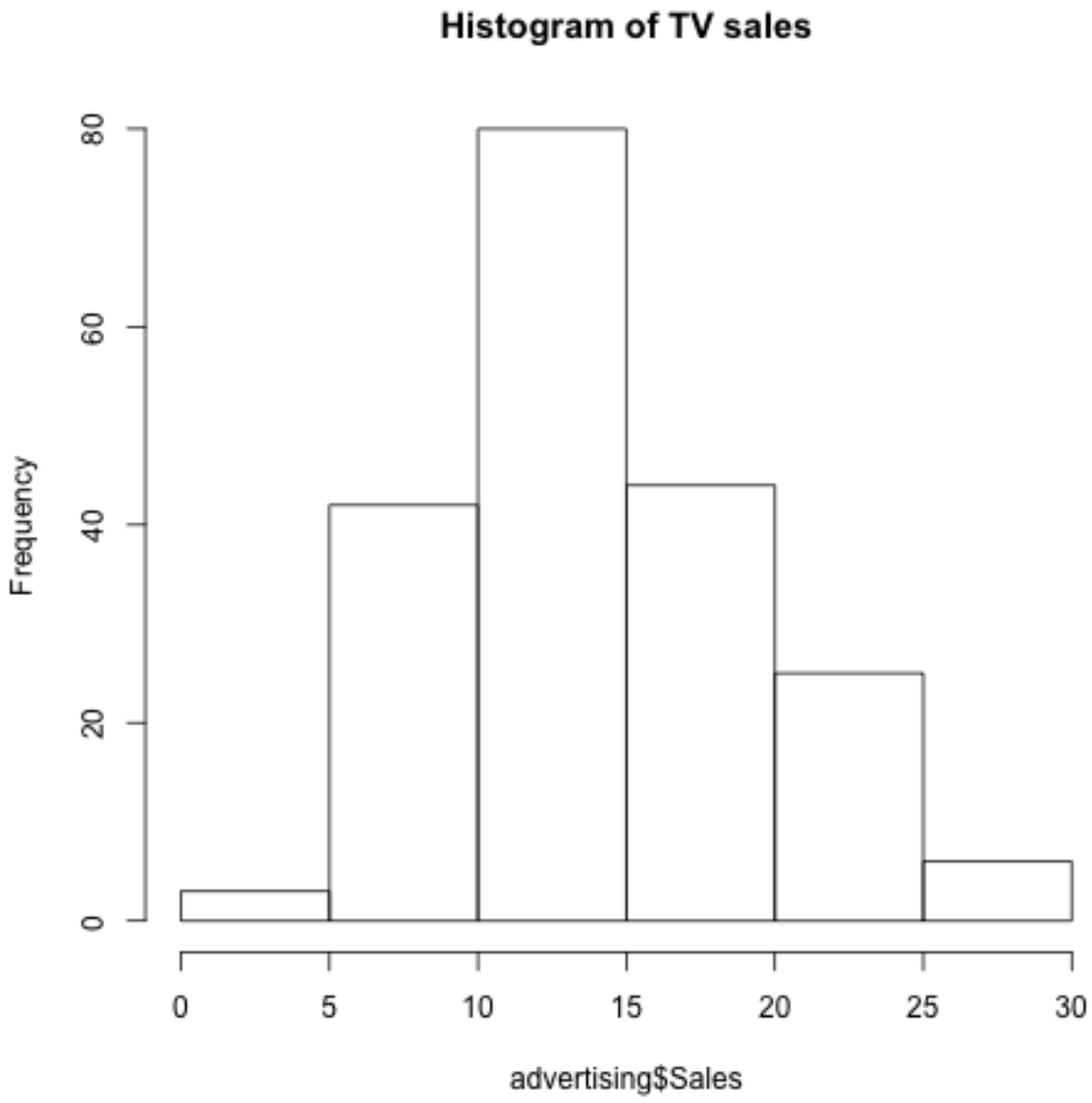

Figure 2: Histogram Sale

As we can see, the frequencies of Sales form a bell shape distribution. Now, let's try to fit a linear model.

Linear Model Results

We compute the regression coefficients in Table 2 below:

From the **Table2**, we can see that the slope of our linear model is around 0.0475 and the intercept point at 7.0326.

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

Table 2: Information about Regression Coefficients

This indicates that if TV advertising budgets increase by 1 thousands dollars, Sales will increase by about 0.0475 thousands of units. This also indicates that if TV advertising budget is zero, Sales will still be around 7.0326 thousands of units.

Now let's see if our model is reliable or not.

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

Table 3: Regression Quality Indicies

From the **Table3** above, we can see that the **R Squared**, which is a indicator of correlatin between Sales and TV advertising budgets,is 0.61. I would say 0.61 is not very high, which means it's not good enough.

The **Residual Standard Error**, the average amount that Sales deviate from the true regression line, is 3.26. That means the average deviation of Sales from our predicted value is about 3260 units. 3260 is fairly large since the mean of Sales is only around 14000. So I would say our predicted linear model is not very reliable.

At last, from the scatterplot graph of **Figuer 3**, we can see that the deviation becomes larger and larger as TV advertising budgets increase.

That means our model becomes more unreliable as TV advertising budgets increase.

Conclusion

Our linear regression model is able to make relatively reliable prediction when TV advertising budget is small. The reliability of our prediction decreases as TV advertising budget gets larger. This problem can be possiblely solved by fitting a model from a larger dataset, fitting a different model or removing outliers.

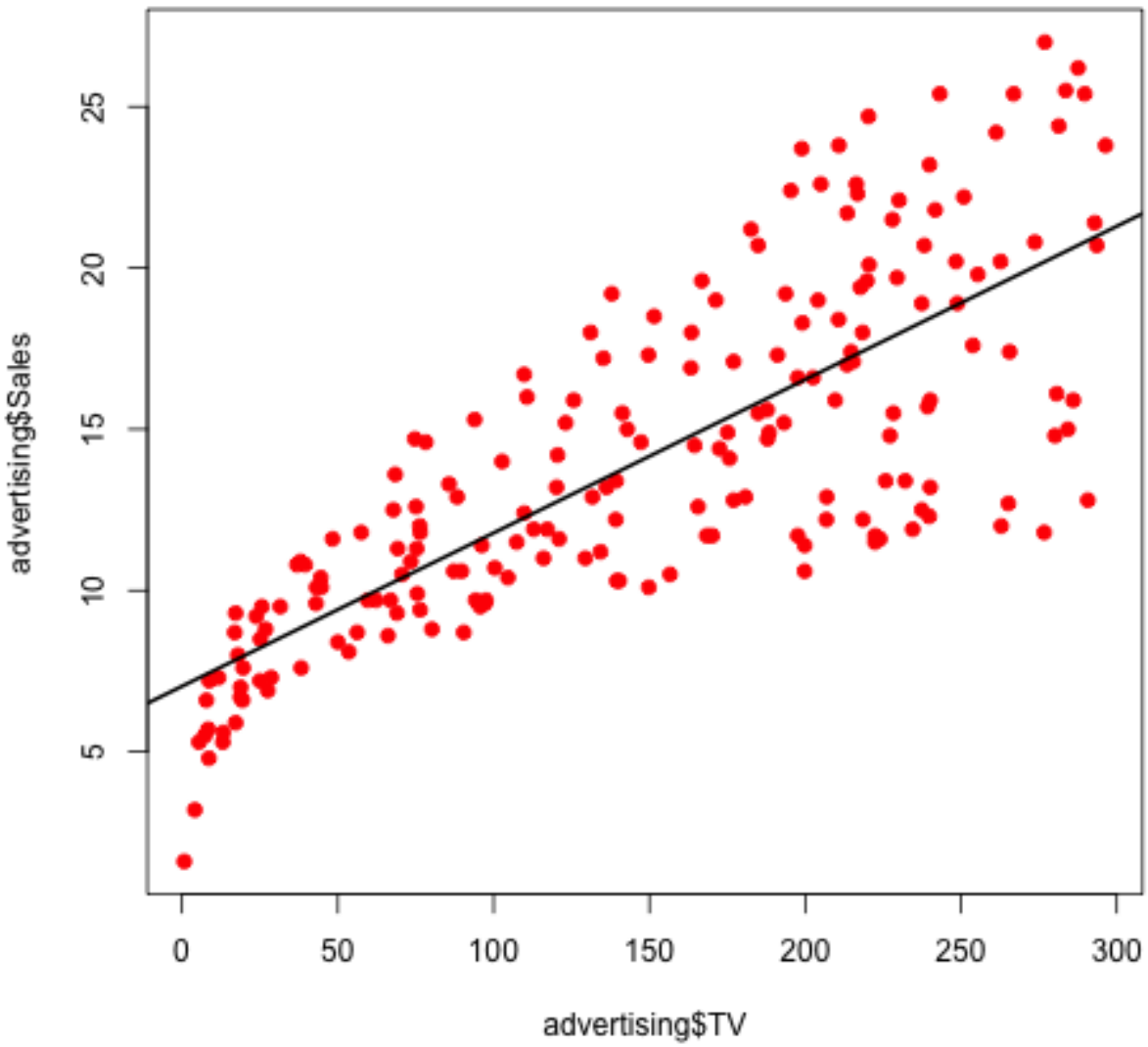

Figure 3: Scatterplot with fitted regression line