Simple Regression Analysis

Steven Chen

Oct 31st, 2016

Abstract

This report reproduces a simple linear regression in the book *An Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. The model uses the explanatary variable, TV advertising budget, to predict sales. For more information, please refer to Chapter 3.1 of the book.

Introduction

The main goal of the analysis is to provide details about whether advertisements through different channels improve sales, and this report will look at television advertisements. We want to learn whether there is a relationship between TV advertisement budget and sales, and if there is we want to describe the relationship with a model, in this case a simple linear model.

Data

The data set we are using, **Advertising.csv**, contains 200 data samples, each sample containing the response variable **Sales** and the explanatary variables **TV, Radio, and Newspaper**. Advertising.csv can be downloaded for free, and credit goes to Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani.

Since we are mainly focused on **TV** and **Sales** in this paper, we have two histograms of the data distribution for the two variables.

As you can see from *Figure 1*, the distribution of the TV data is quite uniform. However, the data of the sales data is more normal.

Methodology

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

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

And to find the values for the two coefficients β_0 and β_1, we fit the linear regression model based on the normal least square criterion. β_0 represents the intercept term and β_1 represents the weight of influence that **TV** plays in predicting the sales. Now one might ask how does the estimation of these parameters work? Since we are using the least square method, the parameters are estimated so that the sum of squares of the predicted sales and the actual sales is minimized.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

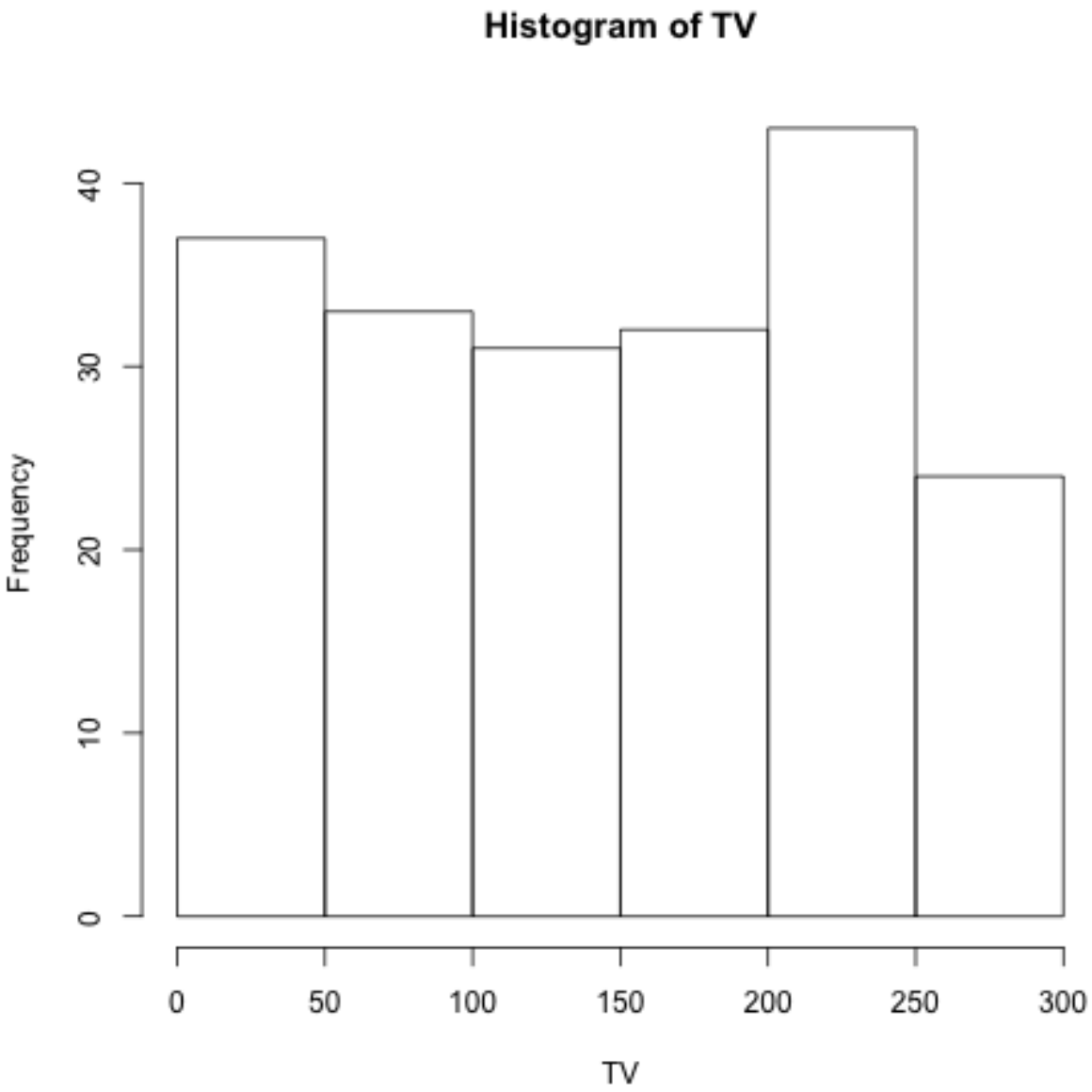

Figure 1: Figure 1: Histogram of TV data

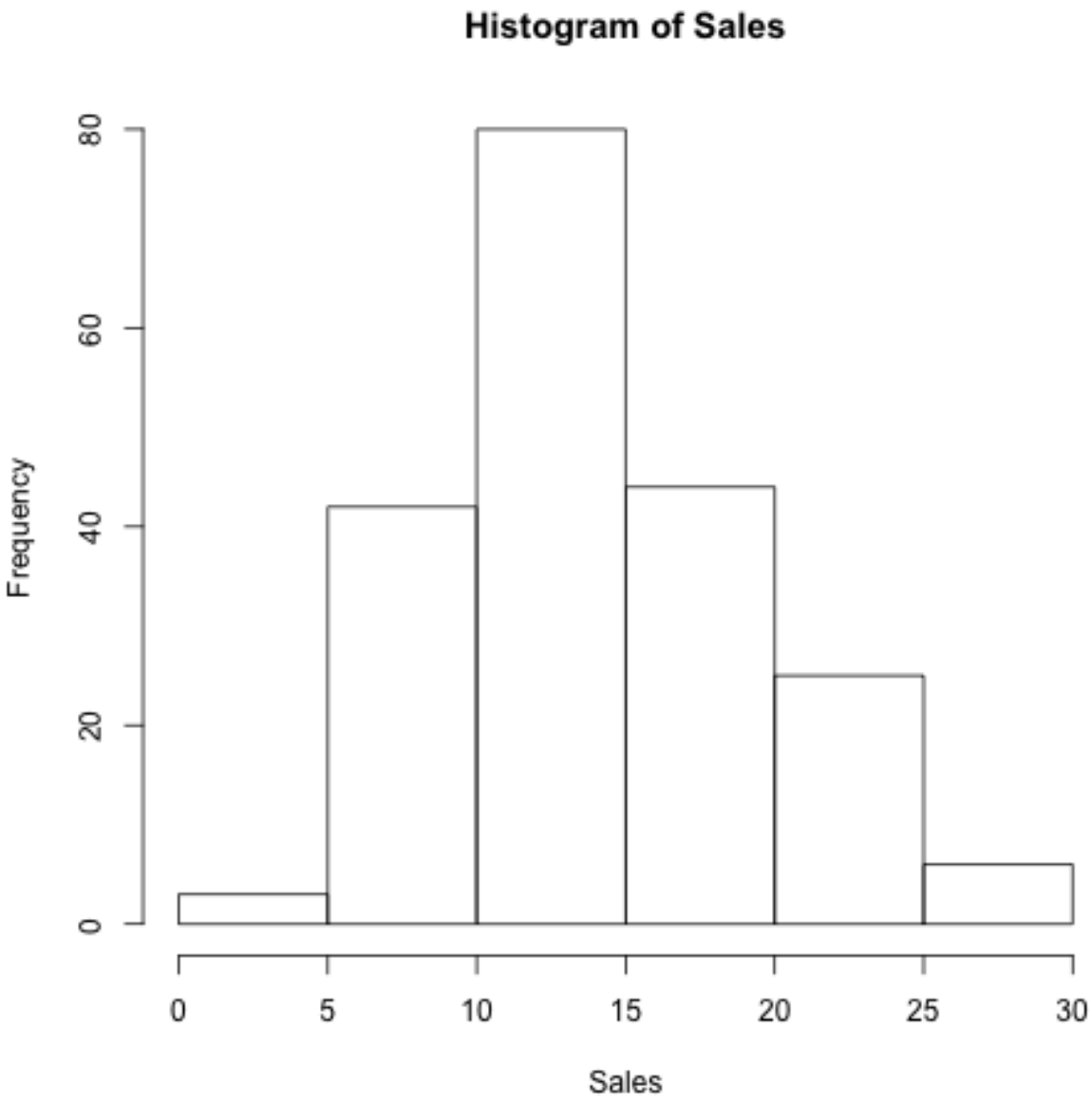

Figure 2: Figure 2: Histogram of Sales data

Results

Since we are trying to find a linear model, it's extremely easy in R to achieve that and get valuable summary feedbacks. Below shows the code to read in the data set and to fit a linear model with TV being the explanatary variable and Sales being the response variable. The tables following that are the summary statistics that the linear model contains:

```
advertising = read.csv("data/Advertising.csv", header = T)
model = lm(Sales ~ TV, data=advertising)
summary(model)
```

	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

Quantity	Value
Residual standard error:	3.259
R-squared:	0.6119
F-statistic:	312.1

After we fit the model, we can plot the data points in a scatter plot as well as the fitted linear model to better visualize the relationship between **TV** and **Sales**.

```
plot(Sales~TV, data=advertising, main="Plot of TV and Sales")
abline(model,col='red')
```

Conclusions

From Figure 3, we can see that the relationship between **TV** and **Sales** is quite linear, and it's also a positively correlated relationship. Since the null hypothesis is that there is no relationship between **TV** and **Sales**, the close to zero p-value of the TV coefficient means that we can reject the null hypothesis. Based on the model, we can conclude that an increase in **TV** advertising budget will lead to increase in **Sales**. To be more precise, an increase in $1000 in **TV** advertising budget will lead to around 47.5 unit increase in **Sales**.

The linaer model we have here is decent, but improvements can be made by increase the number of data samples as well as adding new features into the model that could provide more insight into whether increase in advertising will indeed increase sales.

Figure 3: Figure 3: Scatterplot with fitted regression line