Simple Linear Regression

Bryan Alcorn

October 31, 2016

1 Abstract

We will learn to combine our previous skills and make a simple report from a MakeFile.

Notes from the Professor So far we have been focused on introducing and learning the basic tools typically used in computational reproducible workflows (e.g. bash, git, github, Make, markdown, pandoc, and some text editor). However, we haven't done any statistical data analysis . . . yet.

The purpose of this assignment is to give you the opportunity to start applying the computational toolkit (plus R) to reproduce a simple regression analysis from the book ”An Introduction to Statistical Learning”

2 Introduction

We want to be able to replicate this result, possibly if the data changes or anthing else changes, we can call upon our MakeFile to change everything for us.

The analysis involves carrying out a simple linear regression of ‘TV‘ advertising on ‘Sales‘ of a particular product. The overall idea is to write a report in which you are able to replicate.

3 Data

Along with the description of the Advertising data set bellow by the professor, we also utilize an RData file with the regression results in it and another text file with the statistical summaries of the 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.

Utilizing the function lm (linear model), we get the coefficients for a regression line. Using the method of least squares, we estimate coefficients for the equation $y = mx + b$

$$Sales = m(TV) + b \tag{1}$$

4 Steps takes for the project

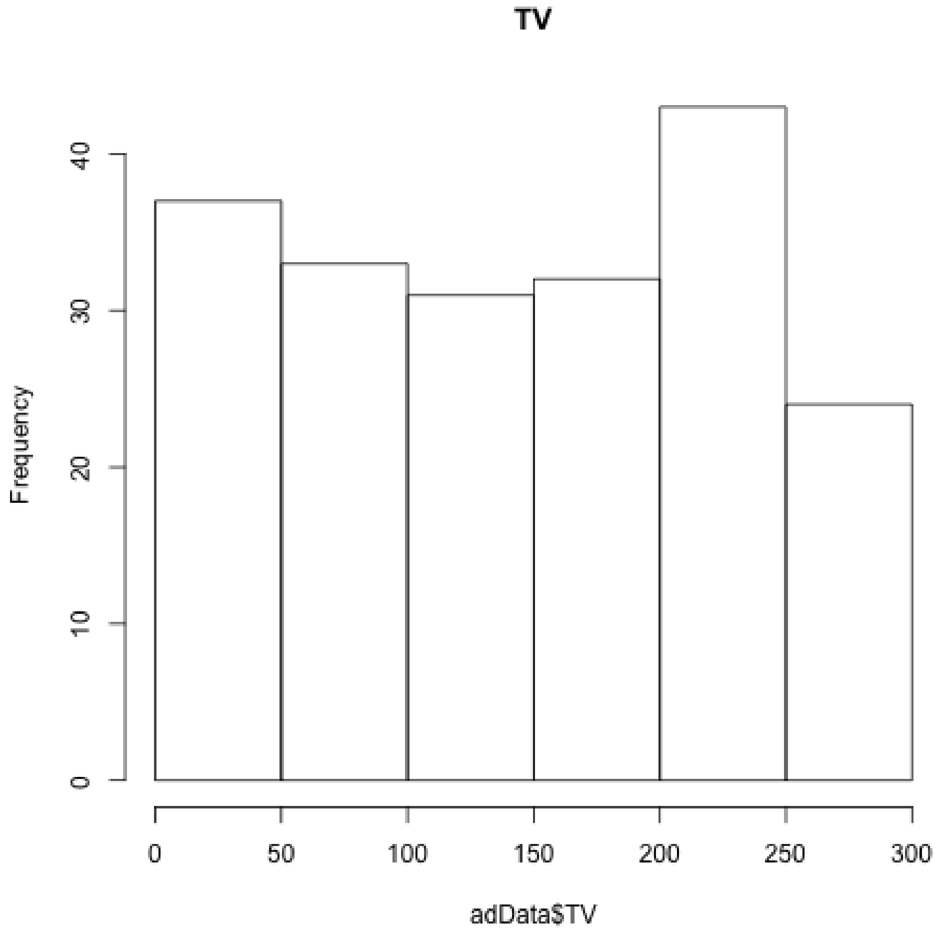
- gather data
- produce the linear model from the data
- export all data and create the images
- Make the make file connecting everything
- Write the report and then run the make file to output a PDF file of the report

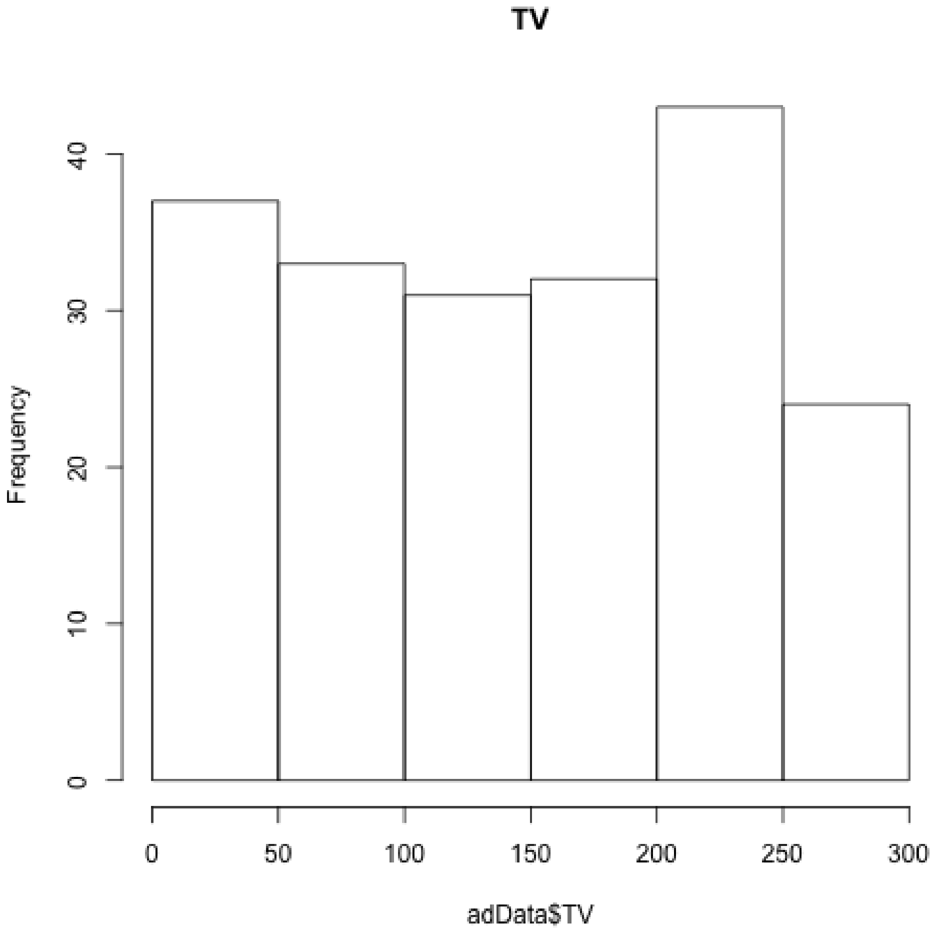
5 Results

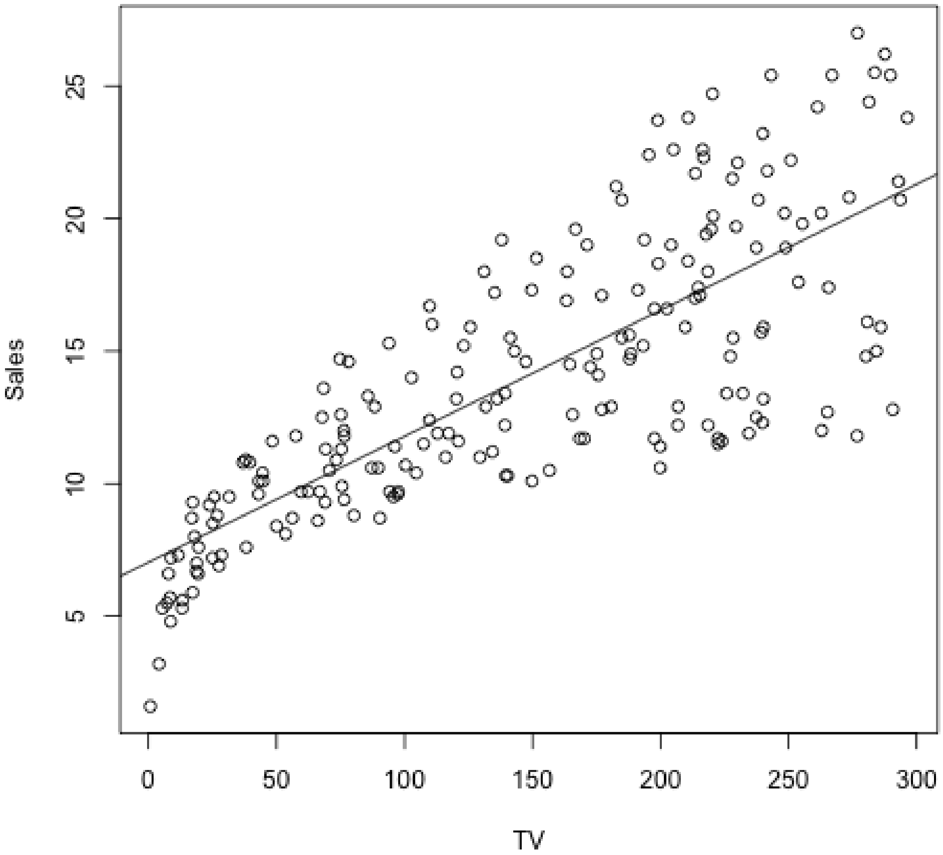
```
> load("../../HW/hw2/data/regression.RData")
> summary = summary(adModel)
> summary

Call:
lm(formula = adData$Sales ~ adData$TV)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.3860 -1.9545 -0.1913  2.0671  7.2124 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
adData$TV   0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,	Adjusted R-squared:  0.6099 
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

The relationships seems to be corrolated. the R-Squared value is above .6 which indicates a relationship, but not a strongly corrolated one.

Bellow are some plots to illustrate the data provided in Advertising.csv


TV
Frequency
0
10
20
30
40
0
50
100
150
200
250
300
adData$TV


TV
Frequency
0
10
20
30
40
0
50
100
150
200
250
300
adData$TV



The data has an interesting trend. The spread increases over time. As the TV variables increases, the data points become less concentrated around the sales data. Also interesting is the large P value for the data. This does not seem to be data produced by chance.

6 Conclusion

To summarize, in this proejct I automated my workflow to gather data and produce the images and the data for a report. Then the RMD file synthesizes that information and uses its dynamic capabilities to include the images and code in the report.

We analyzed the relationship between sales and other variables and there seemed to be a relationship between TV and Sales. There is a high p value $\implies$ results not likely due to chance