Multiple Regression Analysis

Steven Chen

Oct 14th, 2016

Abstract

This report reproduces a multiple 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 3.2 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, radio, and newspaper advertisements. We want to learn whether there is a relationship between these different advertisement budgets and sales, and if there is we want to describe the relationship with a model, in this case a multiple linear model. To load the required data and functions, follow the instructions:

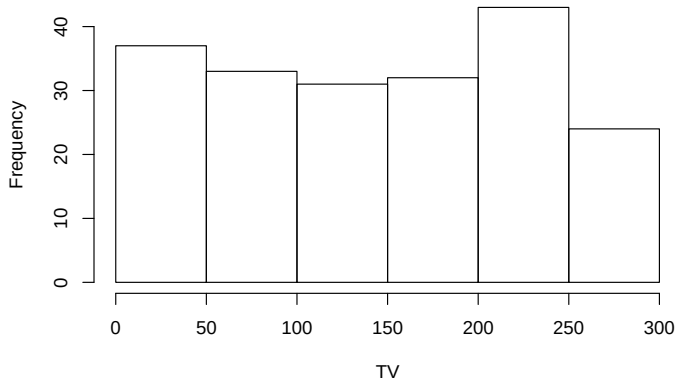
Data

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**, **Radio**, **Newspaper**, and **Sales** in this paper, we should look at the distribution of each variable:

```
hist(advertising$TV, xlab="TV", main="Figure 1. Histogram of TV")
```

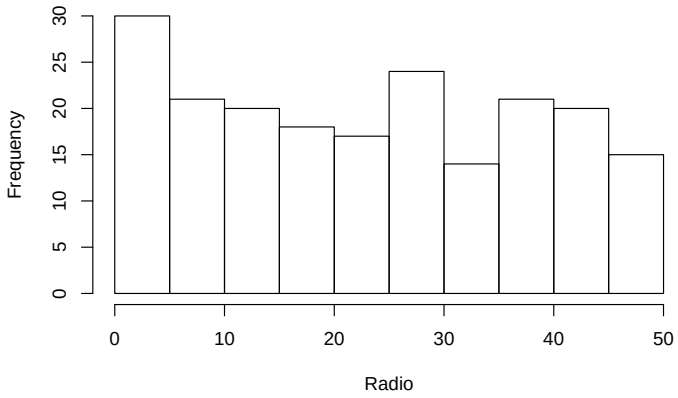
Figure 1. Histogram of TV



As we can see from the histogram, the data for TV advertisement budget looks quite uniform.

```
hist(advertising$Radio, xlab="Radio", main="Figure 2. Histogram of Radio")
```

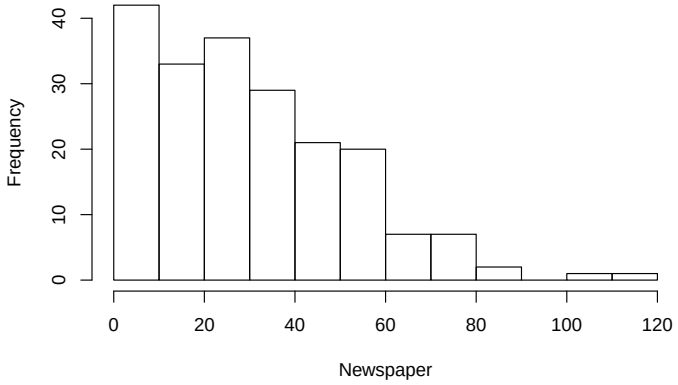
Figure 2. Histogram of Radio



As we can see from the histogram, the data for Radio advertisement budget looks uniform as well, except for the left tail.

```
hist(advertising$Newspaper, xlab="Newspaper", main="Figure 3. Histogram of Newspaper")
```

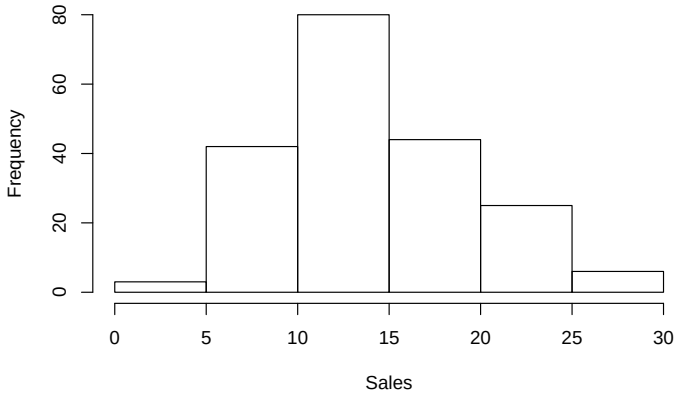
Figure 3. Histogram of Newspaper



As we can see from the histogram, the data for Newspaper advertisement budget looks quite skewed.

```
hist(advertising$Sales, xlab="Sales", main="Figure 4. Histogram of Sales")
```

Figure 4. Histogram of Sales



As we can see from the histogram, the data for Sales looks quite normal.

Methodology

We consider **Sales** vs **(TV | Radio | Newspaper)** in our dataset and try to fit them in a simple linear regression model:

$$Sales = \beta_0 + \beta_1(TV|Radio|Newspaper)$$

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 | Radio | Newspaper)** 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.

Instead of fitting simple linear regression model for each explanatory variable, we can try fitting all of them together in a multiple linear regression model. The model extends to:

$$Sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper$$

Results

Sales ~ (TV | Radio | Newspaper)

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, Newspaper, and Radio being the explanatary variable and Sales being the response variable. The tables following contains information about the coefficients:

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

Table 1: Simple Regression Coefficients of Sales on TV

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.31	0.56	16.54	0.00
Radio	0.20	0.02	9.92	0.00

Table 2: Simple Regression Coefficients of Sales on Radio

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.35	0.62	19.88	0.00
Newspaper	0.05	0.02	3.30	0.00

Table 3: Simple Regression Coefficients of Sales on Newspaper

As we can see, the p-value for TV, Newspaper, and Radio are both very small, thus showing strong statistical significance. Since they all seem to be statistical significant, we could fit all the variables together and see if that improves the model.

Sales ~ TV + Radio + Newspaper

From the table, we can see that TV and Radio both have very low p-values, which means they are statistical significant. However, Newspaper has a high p-value, meaning it might not be a good predictor.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.94	0.31	9.42	0.00
TV	0.05	0.00	32.81	0.00
Radio	0.19	0.01	21.89	0.00
Newspaper	-0.00	0.01	-0.18	0.86

Table 4: Multiple regression of Sales on TV, Radio and Newspaper

We can use other statistics for the model to gain further information about the quality of the model.

	Quantity	Value
1	R2	0.897210638178952
2	RSE	1.68551037341474
3	F-stat	570.270703659094

Table 5: Summary Statistics for model of Sales on TV, Radio and Newspaper

As we can see the R^2 value is very close to 1, meaning that the model is a good fit of the data.

Conclusions

As we can see that the relationship between **Sales** and multiple other advertisement variables is strong, 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 50 unit increase in **Sales**. In addition, $1000 increase in **Radio** will lead to around 190 unit increase in **Sales**. Since **Newspaper** is not significant, we can disregard that variable.

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.