report

Lingjie Qiao

10/14/2016

Abstract

This report aims to reproduce the main results displayed in **section 3.2: Multiple Linear Regression** of the book ***An Introduction to Statistical Learning*** and perform simple linear regression analysis on the data set `Advertising`.

Introduction

According to the book, the overall goal is to provide advice on how to improve sales of the particular product. 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. Rather than comparing variables separately, we fit a multiple linear regression model, as discussed in the methodology part to analyze such association.

Data

We download the data set `Advertising`, which is provided by the author of the book. This data set has four variables: `TV`, `Radio`, `Newspaper`, and `Sales`. It 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.

Methology

In this paper, we mainly consider the relatinoship between Sales versus **TV**, **Radio** and **Newspaper**. In order to explore this multiple variable relationship, we use a multiple linear model and regress `sales` onto `TV`, `Radio`, `Newspaper` by fitting the model:

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

Mathematically, β_0 represents the intercept and β_1 to β_3 represents the slope terms in the linear model. With this linear model, we estimate the coefficients by minimizing the least squares criterion, which is minimizing the sum of squared errors.

Results

With the least square estimators, we compute the regression coefficients:

Table 1: Information about Regression Coefficients

% latex table generated in R 3.2.3 by xtable 1.8-2 package % Fri Oct 14 19:03:34 2016

Table 1: For the Advertising data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.

More information about the least squares model is given in the table below:

	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 1: Regression Coefficients

Table 2: Correlation Matrix

% latex table generated in R 3.2.3 by xtable 1.8-2 package % Fri Oct 14 19:03:34 2016

	TV	Radio	Newspaper	Sales
TV	1.00	0.05	0.06	0.78
Radio	0.05	1.00	0.35	0.58
Newspaper	0.06	0.35	1.00	0.23
Sales	0.78	0.58	0.23	1.00

Table 2: Correlations Matrix

Table 2 indicates correlation matrix for TV, radio, newspaper, and sales for the Advertising data.

Table 3: Regression Quality Indices

```
#source('../code/functions/regression-functions.R')
#RSE = residual_std_error(lm_ad)
#R_square = r_squared(lm_ad)
#f_statistic = f_statistic(lm_ad)

#df = data.frame(Quantity= c("Residual Standard Error", "R-squared", "F-statistic"), Value = c(RSE, R_s
#print(xtable(df, caption = 'Regression Quality Indices'), comment = FALSE, include.rownames=FALSE, cap
```

Here are some sample images relating Advertising dataset

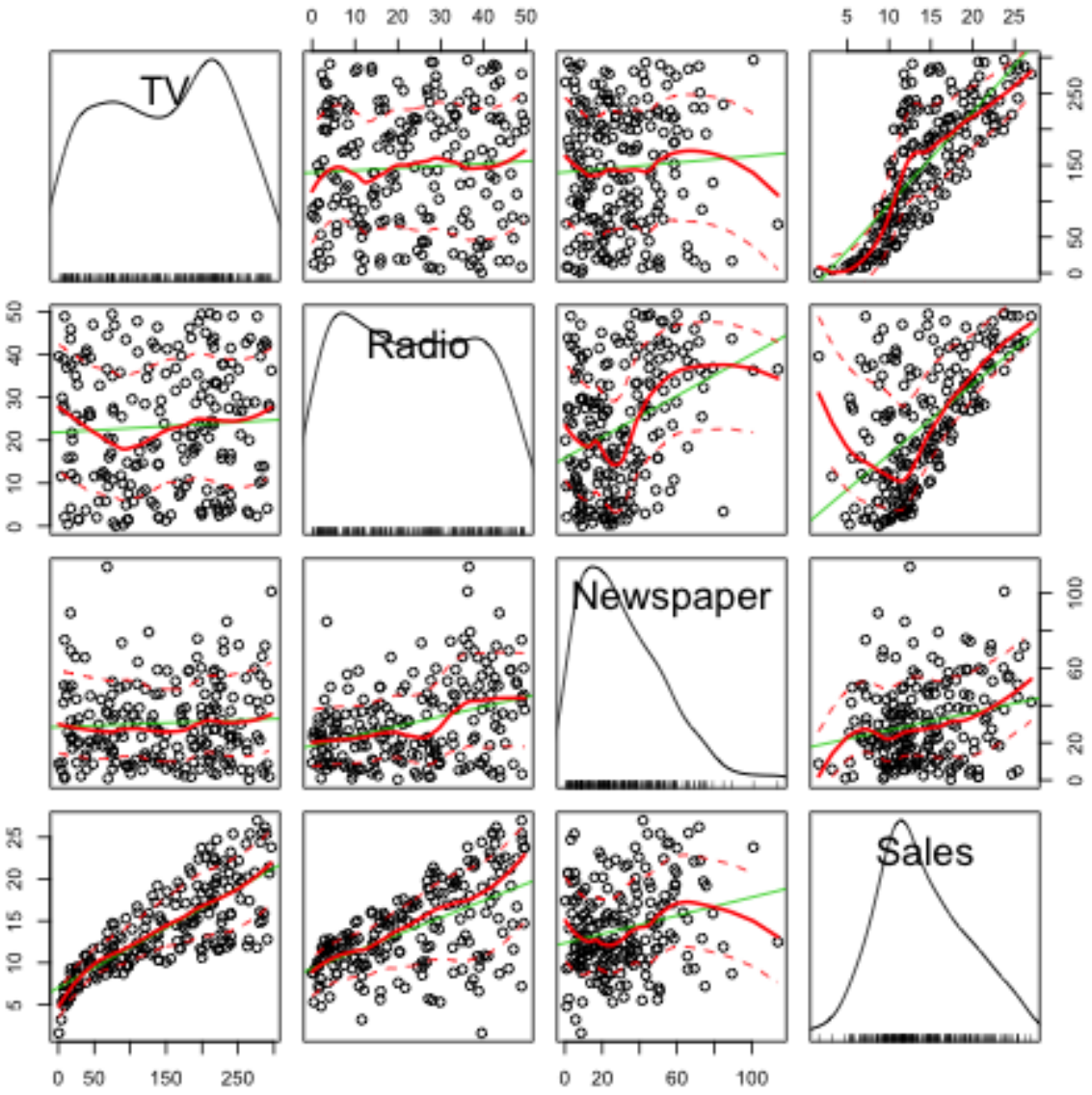

Figure 1: Scatterplot Matrix

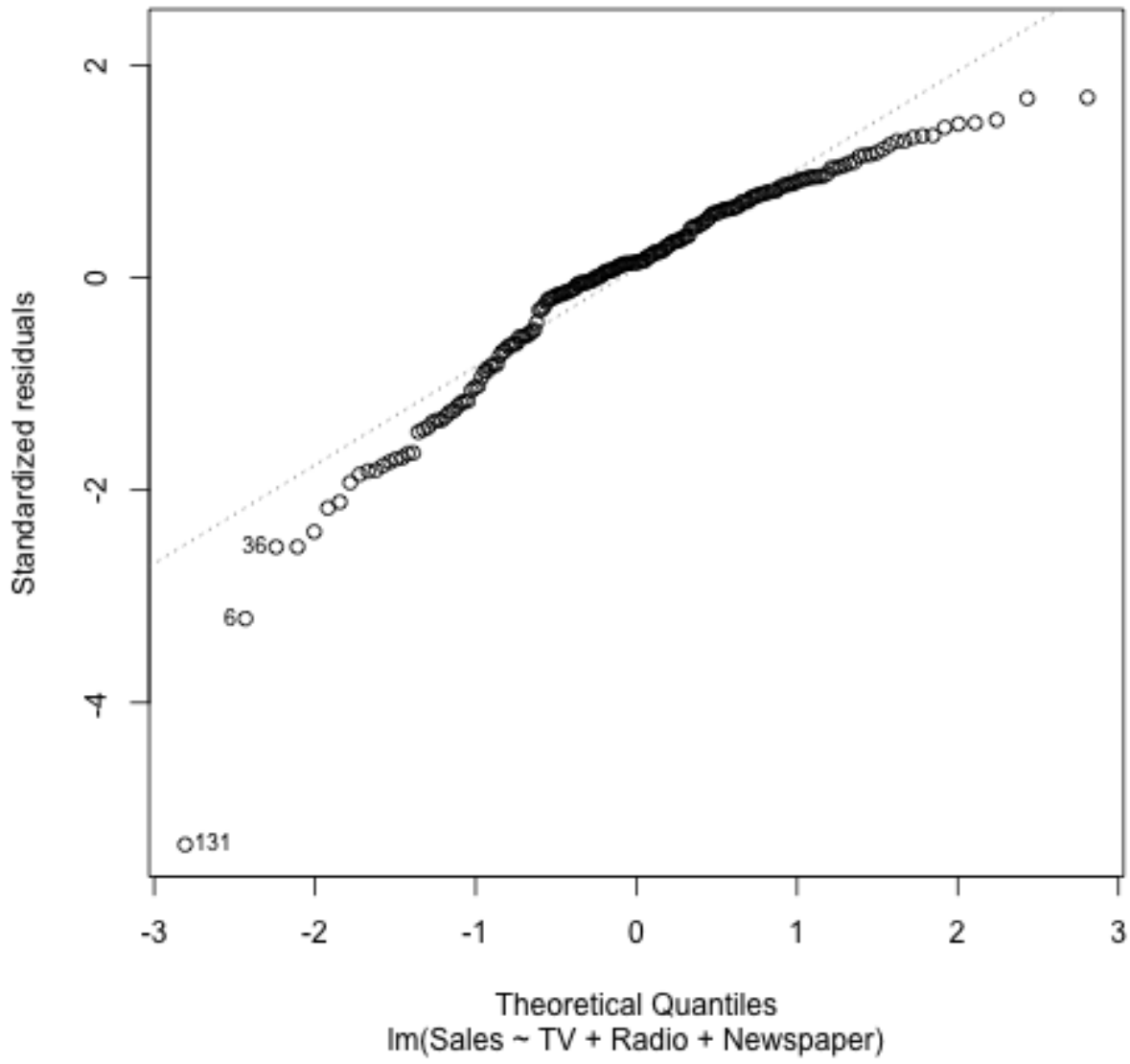

Figure 2: Normal Q-Q plot

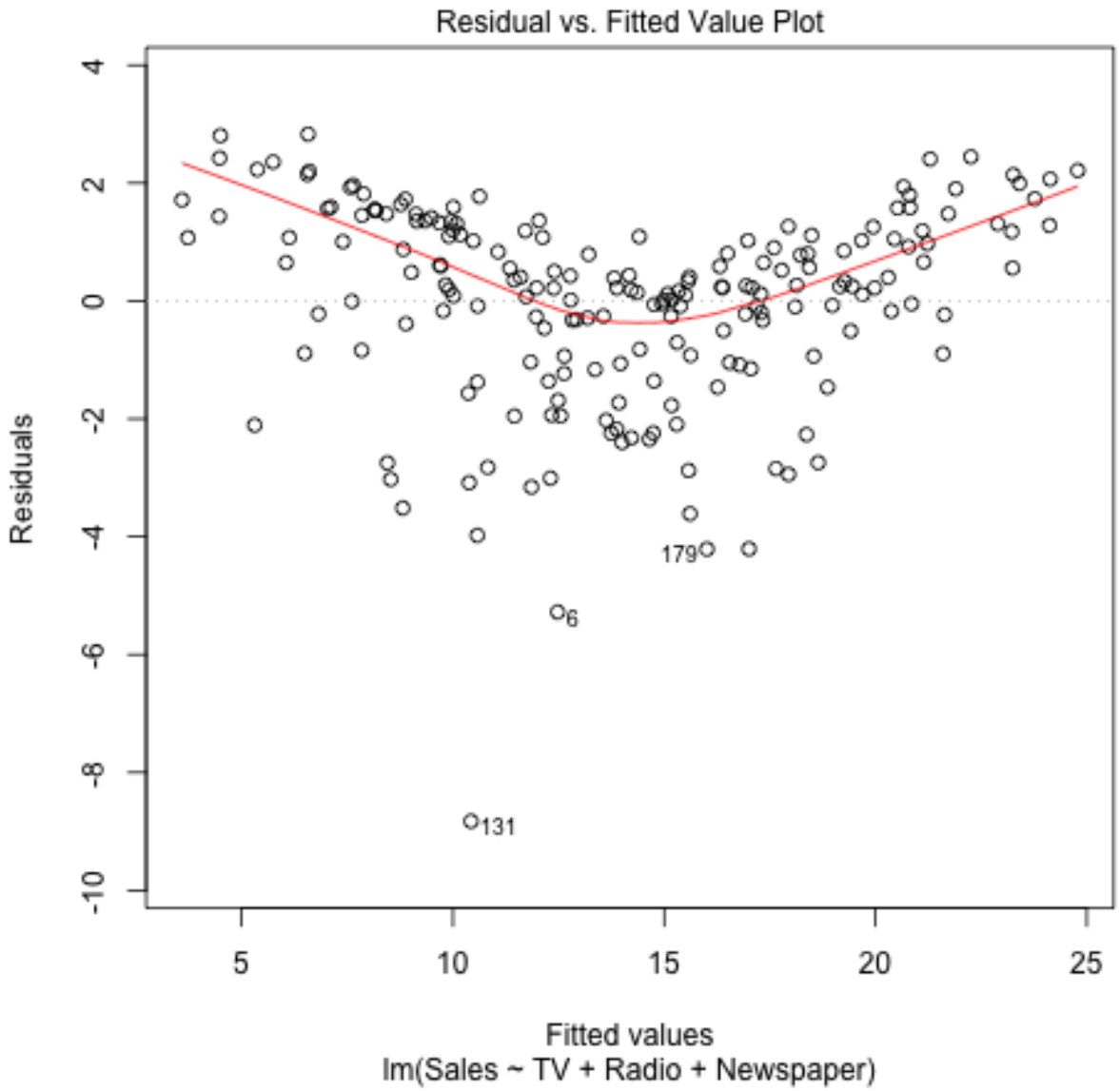

Figure 2: Residual plot

Conclusions

To conclude, I explored the linear relationship between TV, Radio and Newspapers versus Sales, fitting a multiple linear regression model upon the advertising data to understand the information hidden in the data. From the reproduced graph we can see the same results as produced in the book, namely "a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot." This project helps us to fully understand the multiple linear regression model, its mathematical interpretation, and all the data retrieved from the R fitted linear model. It also gives us great insights in the reproducible project, and in specific, how important it is to be able to reproduce other people's work.