Report - Multiple Linear Regression

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

First, for the sake of comparison, we want to first put out the coefficients estimates of simple regression models: `Sales` on `TV`, `Sales` on `Radio`, and `Sales` on `Newspaper`.

Table 1: Simple Linear Regression on TV and Sales

Table 2: Simple Linear Regression on Radio and Sales

Table 1: Simple Linear Regression on TV and Sales

	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 2: Simple Linear Regression on Radio and Sales

	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 3: Simple Linear Regression on Newspaper and Sales

Now, with the least square estimators, we compute the multiple linear regression coefficients:

Table 4: Information about Regression Coefficients

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.

Some reflections we can make on this table is looking at the estimates. It seems that `TV` and `Radio` both have positive impact on the sales, while `Newspaper` has negative impact. When we compare TV and Radio, since $0.05 < 0.19$, `Radio` seems to have a higher influence on sales. This gives us a basic understanding of the association between sales with media variables - if there is a positive or negative impact.

Through looking at p value, we see that `TV`, `Radio` and `Newspaper` all have very small p-value (near 0), which is supposed to reject the null hypothesis. Since the null hypothesis states that the coefficient is 0, it means that all three variables actualy have an impact on the sales data. Therefore, we can reach the conclusion that all predictors to some extent help to explain the response, indicateing the usefulness and responsiveness of costumers on sales.

From my perspective, the linear model to some extent does not fit the data well. When we take a closer look at the individual scatterplots, we could hardly see strong linear relationship as most of the points, esepcially in newspaper, are just randomly spreading out. Therefore, even though we fit correctly the linear model, the predictions will not be as accurate as we imagine it to be.

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

Table 5: Correlation Matrix

Table 5 indicates correlation matrix for TV, radio, newspaper, and sales for the Advertising data.

Table 6: Regression Quality Indices

Here are some sample images relating Advertising dataset

Table 3: Simple Linear Regression on Newspaper and Sales

	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 4: Multiple Linear Regression Coefficients

	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

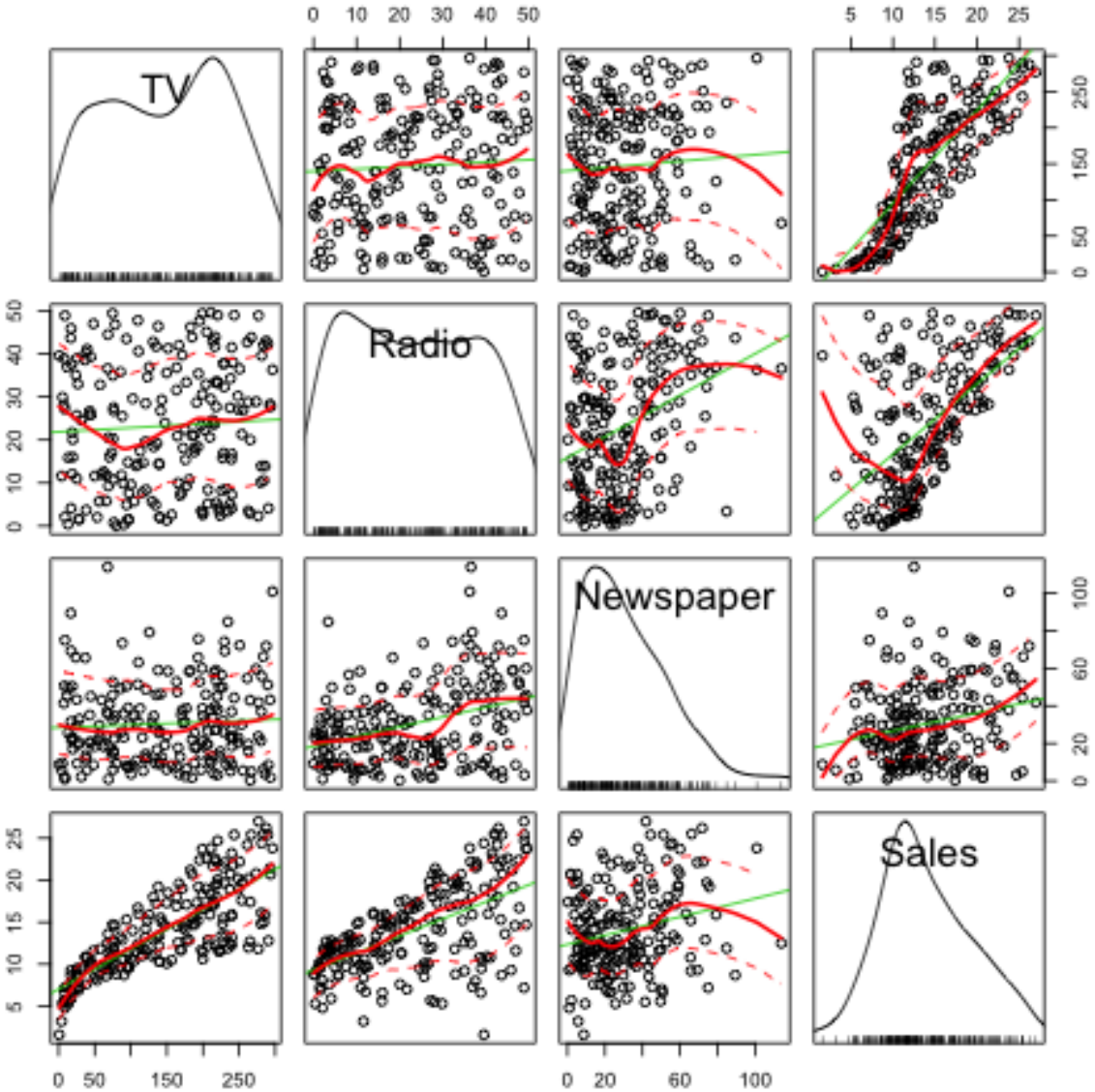
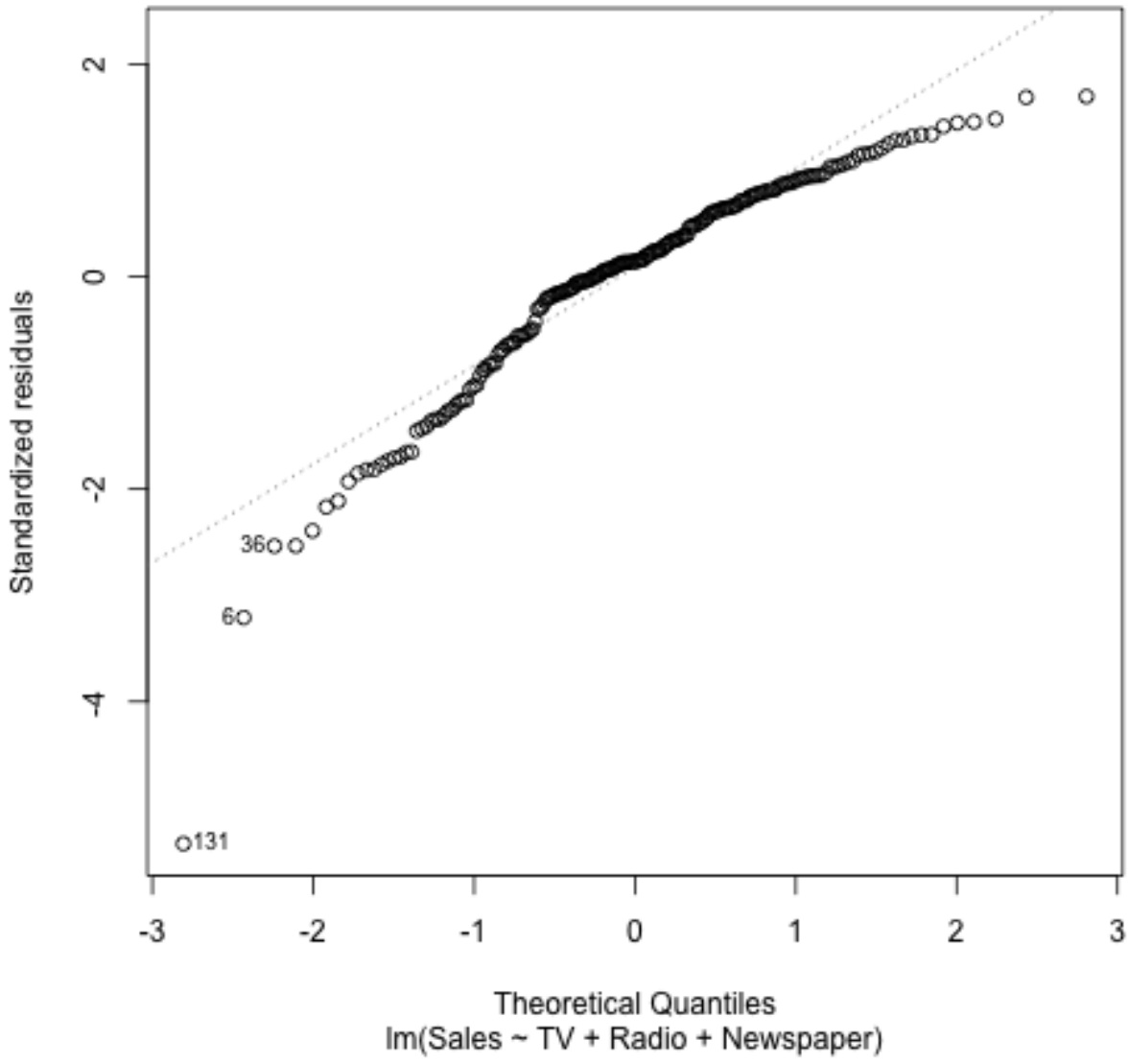


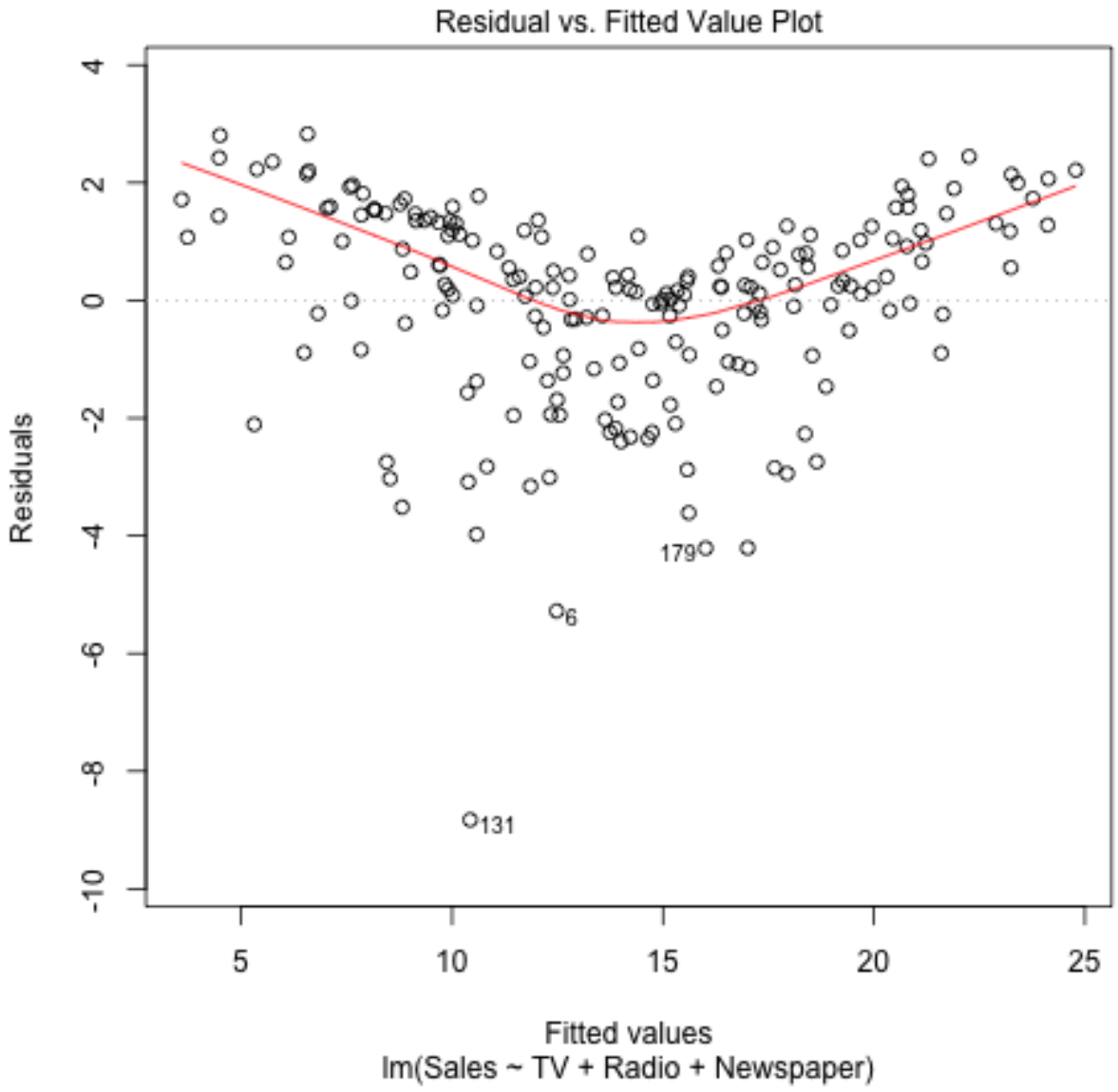
Table 5: Correlations Matrix

	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 6: Regression Quality Indices

	Quantity	Value
1	Residual Standard Error	1.69
2	R-squared	0.90
3	F-statistic	570.27





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.