Multiple Regression Analysis

Kevin Liao

10/14/2016

Abstract

This homework is to reproduce the analysis from Section 3.2 (pages 71 to 82), from the book “An Introduction to Statistical Learning” (by James et al). It includues multiple linear regressin with the predictor variables TV, Radio, Newspaper and the response variable Sales.

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

The dataset **Advertising.csv** comes from *“http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv* It consists for TV, Radio, Newspaper and Sales columns. The structure of the columns are stored in numeric vectors.

Methodology

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

Let’s start with looking at three simple linear regression about how sales is related to each individual feature.

TV Advertisement and Sales

This table provides details of the least squares model for the regression of number of units sold on TV advertising budget for the Advertising data. We can conclude that the two coefficients are not zero.

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

Radio Advertisement and Sales

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

This table provides details of the least squares model for the regression of number of units sold on Radio advertising budget for the Advertising data. We can conclude that the two coefficients are not zero.

Newspaper Advertisement

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

This table provides details of the least squares model for the regression of number of units sold on Newspaper advertising budget for the Advertising data. We can conclude that the two coefficients are not zero.

All Advertisements

Now let's look at a multiple linear regression. We can find out the changes in sales based of these three advertisements and check if there is a correlation between the advertisements. The multiple linear regression with 3 predictors equations is the following:

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

Here is the table stat:

Table 4: Multiple Linear Regression

	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

The simple and multiple regression coefficients can be quite different. This difference stems from the fact that in the simple regression case, the slope term represents the average effect of a $1,000 increase in newspaper advertising, ignoring other predictors such as TV and radio. In contrast, in the multiple regression setting, the coefficient for newspaper represents the average effect of increasing newspaper spending by $1,000 while holding TV and radio fixed. But do all of these statistics make sense? Let's look at correlation between each feature.

Correlation Matrix

Table 5: Correlations Matrix

	X	TV	Radio	Newspaper	Sales
X	1.00	0.02	-0.11	-0.15	-0.05
TV	0.02	1.00	0.05	0.06	0.78
Radio	-0.11	0.05	1.00	0.35	0.58
Newspaper	-0.15	0.06	0.35	1.00	0.23
Sales	-0.05	0.78	0.58	0.23	1.00

This table indicates correlation matrix for TV, radio, newspaper, and sales for the Advertising data. This correlation matrix explains a lot. From previous three each simple linear models, we learned that all of the predictors are useful in predicting the response. Now again looking at the correlation, we make verify that there is clear strong correlation between sales and TV, Radio. Although we also found out Newspaper has the weakest relation with sales, yet it is some what related with Radio.

Next, let's examine the quality of fitness by analyzing key statistics

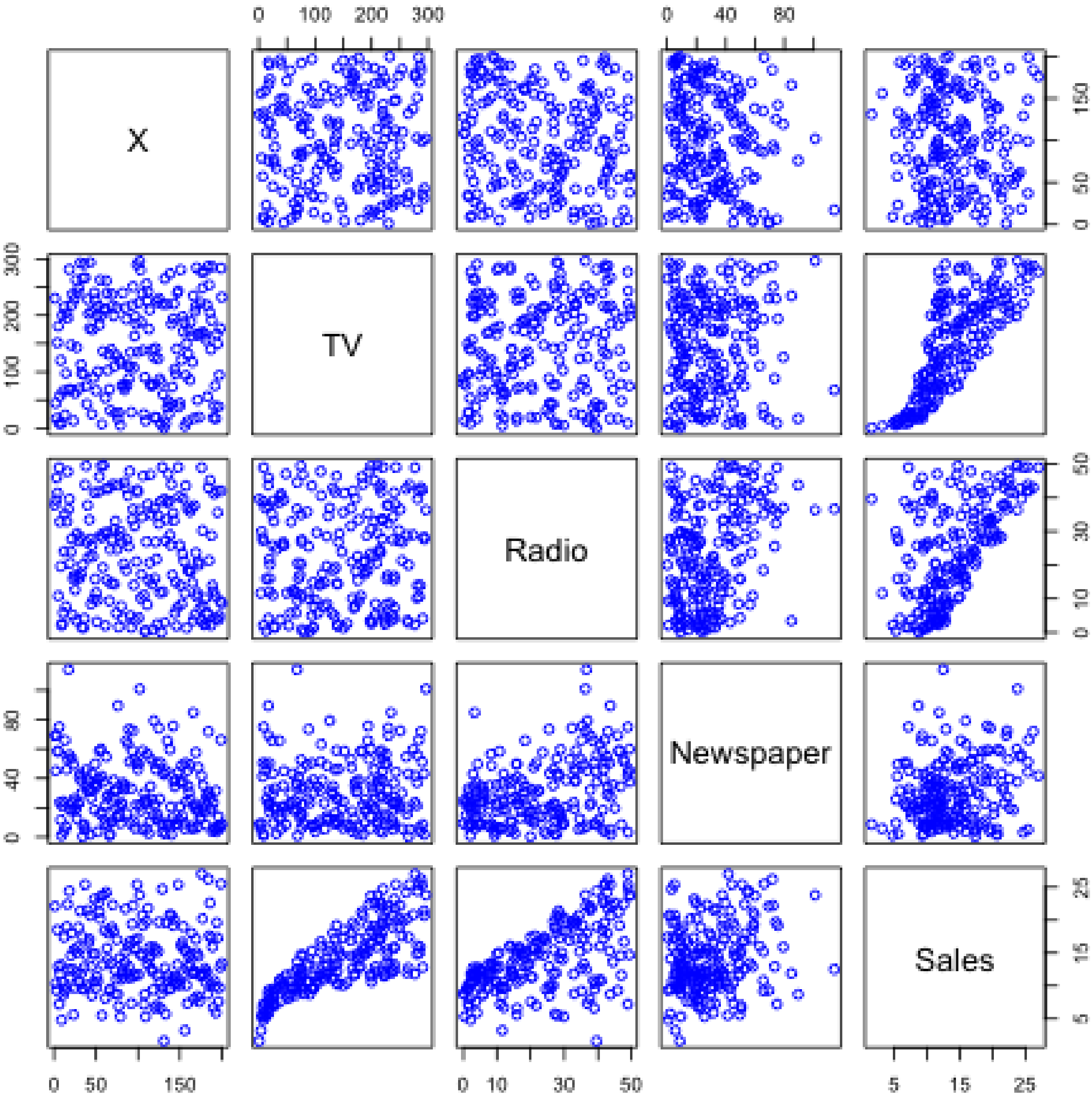
Regression Quality Indices

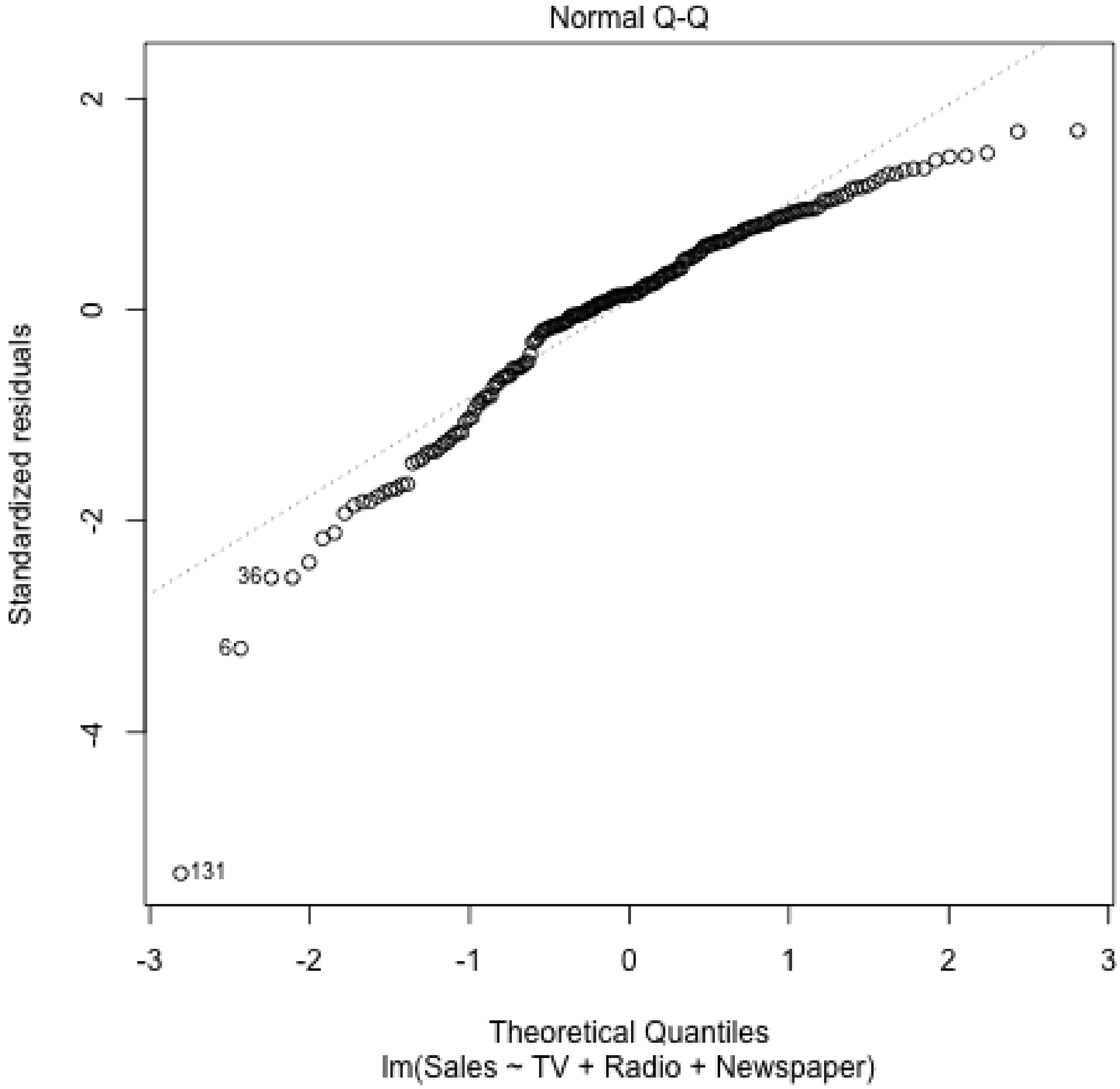
Table 6: Regression Quality Indices

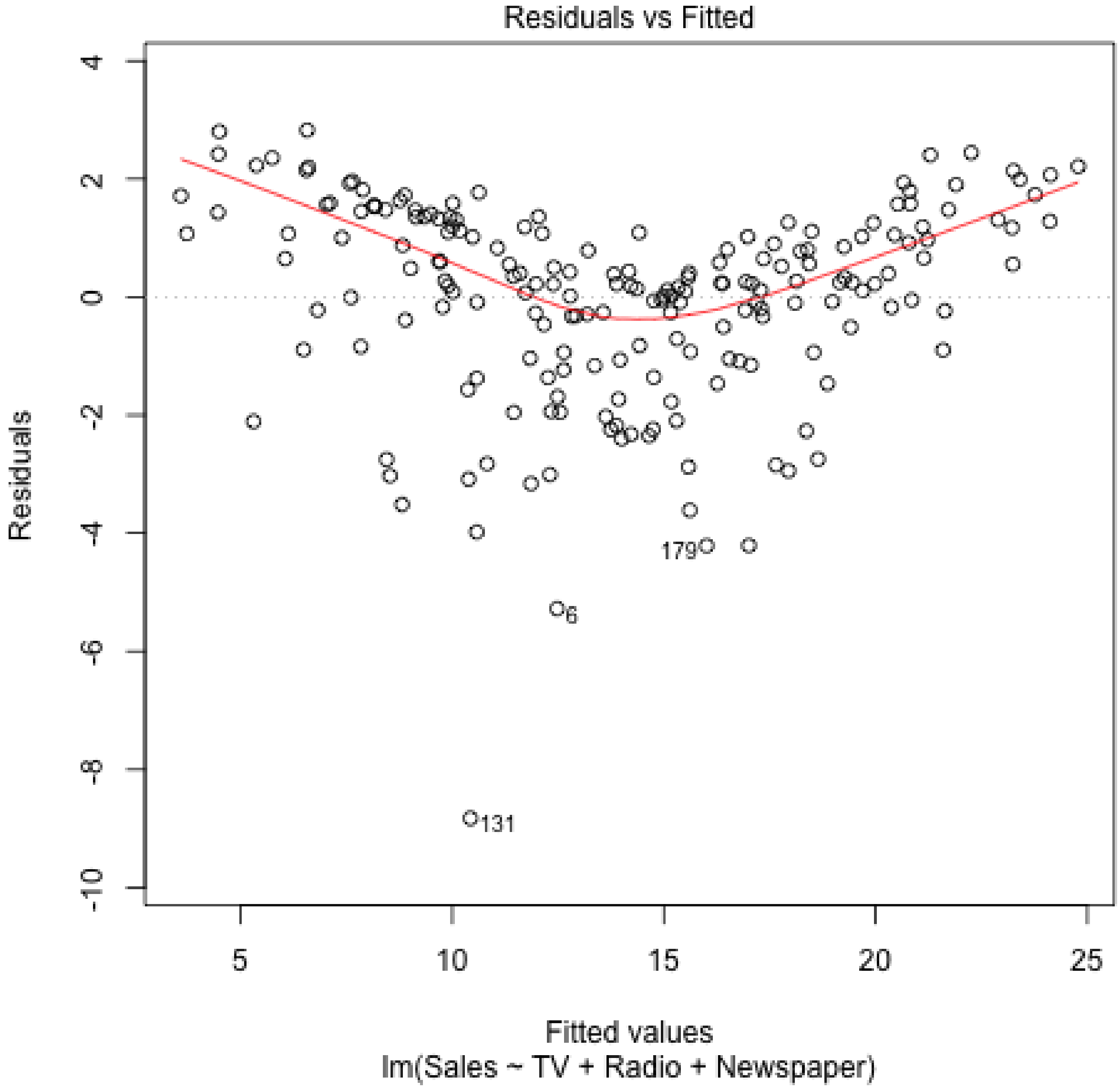
	Quantity	Value
1	Residual Standard Error	1.69
2	R-squared	0.90
3	F-statistic	570.27

From this table, we can first examined R squared, which is .90. This means the data is a good fit to the regression line. The F-statistics is high, which mean at least one advertisement has a correlation with Sales. Further, with conclusion from the correlation matrix, although Newspaper may not be a good predictor, yet TV and Radio are the good predictors. And they fit the data pretty well based on the statistics we seen here.

Here are some sample images relating Advertising dataset


X
TV
Radio
Newspaper
Sales
0
100
200
300
0
40
80
0
50
150
0
10
30
50
5
15
25
0
100
200
300
0
10
30
50
0
40
80
5
15
25
0
50
150


Normal Q-Q
Standardized residuals
2
0
-2
-4
-3
-2
-1
0
1
2
3
36
6
131
Theoretical Quantiles
lm(Sales ~ TV + Radio + Newspaper)



Conclusions

We have learned the simple linear relationship between TV, Radio and Newspapers versus Sales. Now we have fitted a multiple linear regression model upon all of the advertising data so that we could get more insight about the information hidden behind the data. From the reproduced statistics and charts, we have seen the same results as produced in the book. Quoted, "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 better analyze the multiple linear regression model, and what are some mathematical interpretations. It also gives us great insights in the reproducible project. And this project highlights the significance of reproducibility.