Homework3

Leanne Lee

October 12, 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

Given the 3 predictor variables of TV, Radio and Newspaper, we need to find out what is the relationship between the predictor variables and response variable of Sales. The main goal is to find out which type of advertisement is more effective to increase the sales. Therefore, we break down the linear regression from comparing a single variable to multiple variables. By reproducing the result of the regressions, the marketing team can determine which type of advertising should they invest on.

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

For simple linear regression:

$$Sales = \beta_0 + \beta_1 * TV$$

Simple Linear regression is useful when predicting a response based on one single predictor variable. For multiple linear regression:

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

Multiple Linear regression is useful for Advertising data becasue it can determined the relationship between sales and the three types of advertising. We can compare which type of advertisement is more effective and has a stronger association with sales.

Results

```r
library(xtable)
```

```
## Warning: package 'xtable' was built under R version 3.2.3
```

```r
library(png)
library(grid)
load('../data/regression.RData')
load('../data/correlation-matrix.RData')
ad <- read.csv("../data/Advertising.csv")
source("../code/functions/regression-functions.R")
```

Let's compare the coefficients of each single linear regression.

TV Advertisement

	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: TV Advertisement Linear Regression

With $1000 increase in TV advertisement, there will be an increase in sales by 50 units.

Radio Advertisement

	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: Radio Advertisement Linear Regression

With $1000 increase in Radio advertisement, there will be an increase in sales by 200 units.

Newspaper Advertisement

	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: Newspaper Advertisement Linear Regression

With $1000 increase in Newspaper advertisement, there will be an increase in sales by 50 units.

All Advertisements

With 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:

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

By using multiple linear regression, we can see how other predictors changed the sales. From Table 4, we can see that there will be approximately 50 units increase in sales with $1000 budget. Radio still have a better effect of increase in 190 units in sales. However, newspaper has a decrease in sales. Thus, we can concluded

	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: All Advertisements Linear Regression

that radio advertisement is more effective, while newspaper plays a less importatnt role in terms of affect sales. Both TV and radio have low p-values, which means there is a relationship between sales with TV and radio. Newspaper has 0.86 p-value, which means there is no relationship between Newspaper and Sales.

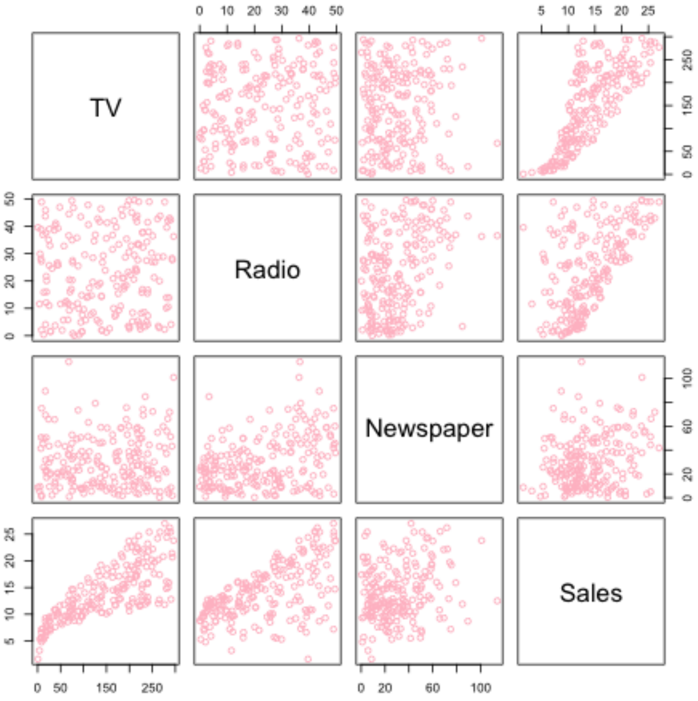
Correlation Matrix

	regstat	name
1	RSS	556.83
2	TSS	5417.15
3	R-square	0.90
4	F-Statistics	570.27
5	Residual Standard Error	1.69

Table 5: Regression Statistics

From Table 5, 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.

Correlation Matrix Graph



In the fourth column, it clearly show that TV and Radio have correlation with Sales. But it doesn't show a correlation between Newspaper and Sales.

Conclusions

From the first three tables of single linear regression model, we find that TV and newspaper advertissment have approximately the same outcome for sales. The major difference is radio advertisement and the marketing team should consider spend more budget on it. By examining the high f-statistics, we can tell that at least one of the predictors useful in predicting the response. The multiple regression model show us that TV and Radio are useful to explain the increase of sales. From the coefficients in the multiple linear regression model, we can see that Radio has the highest effect to increase sales. Newspaper has a negative coefficient and high p-value, which mean there is no relationship between newspaper and sales. The high R squares shows that the model is a good fit to the data; thus, the prediction will be quite accurate.