Multiple Linear Regression Analysis

Andrea Widjaja

10/14/16

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

Stat 159 Homework 2

Abstract

This paper consists of the reproduced results from *Section 3.2 Multiple Linear Regression* from the book **An Introduction to Statistical Learning** *by Gareth James, Daniel Witten, Trevor Hastie, and Robert Tibshirani.* This paper will discuss and examine topics of multiple linear regression analysis such as estimating the coefficients, finding a relationship between response and predictors, deciding on important variables, model fit, and predictions.

Introduction

Simpe linear regression is a useful tool to predict a response on the basis of a single predictor variable. Simple linear regressions are usually in the form

$$Y = \beta_0 + \beta_1 X$$

When taking into account to error, the simple linear regression will be:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

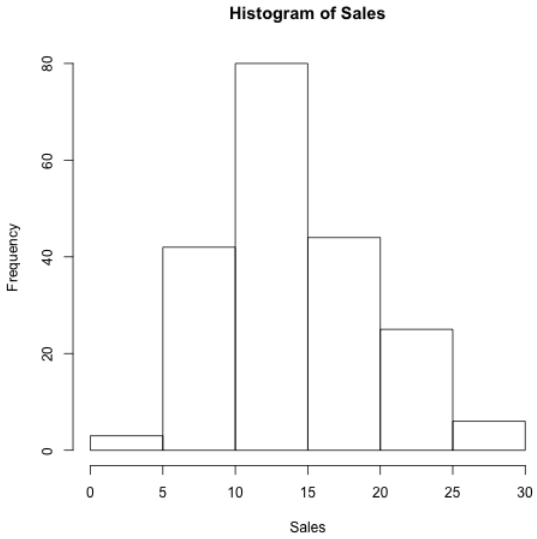
The single predictor variable X are independet variables, whereas the response variable Y are dependent variables. This method of simple linear regression is a very simple approach to predict a quantitative response. In reality, however, there are often more than one predictor. If we approach this problem by fitting a seperate simple linear regression model for each predictor, the estimates generated might be misleading and inaccurate. This is why it better to utilize Multiple Linear Regression.

Multiple Linear Regression is a predictive analysis tool that can interpret relationships between one dependent variable and several independent variables. Hence, it gives us the ability to predict the value of a variable based on the value of several variables.

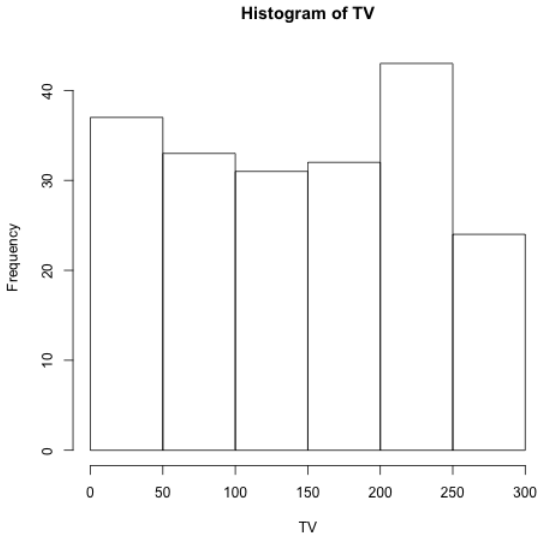
Data

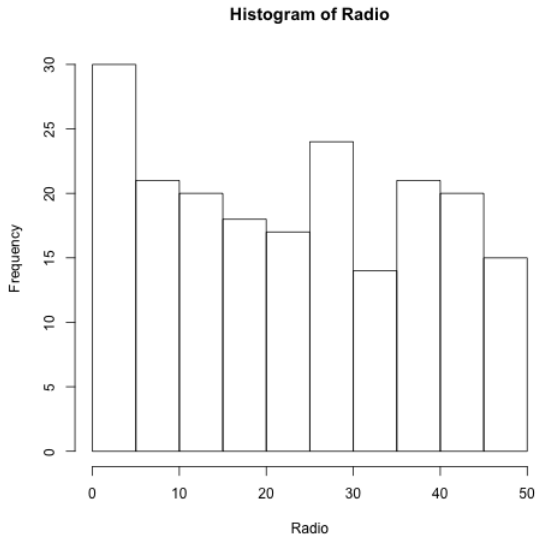
The data set given is *Advertising Data Set*, downloaded from `http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv` called **Advertising.csv**. This data set contains the sales of a product (in thousands of units) across 200 different markets and the advertising budget (in thousands of dollars) for the product in each market. The advertising data contains budgets for three different advertising medias: TV, radio and newspaper. It has 200 records and 5 variables(index, TV, Radio, Newspaper, Sales). Each record contains the information of advertising budget for TV, Radio, and Newspaper, including the sales progress for each market.

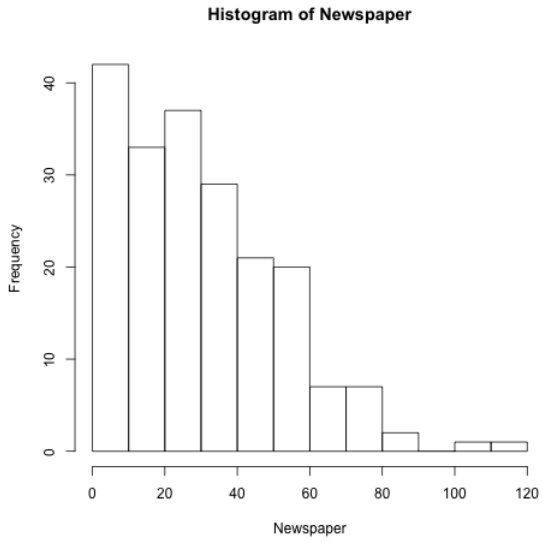
To give a rough picture on what data we are dealing with, these are the histograms for Sales and the



advertising medias TV, Radio and Newspaper.







Methodology

Suppose we know from above that a simple linear regression, taking into account for error, takes the form

$$Y = \beta_0 + \beta_1 X + \epsilon$$

We know that multiple regression have several predictor X variables. Say that we have n numbers of predictors, then the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n + \epsilon$$

where X_i represents the ith predictor, β_i indicates the correlation, β_0 as the intersection, and ϵ as the error. To make it more understandable, β_i represents the average increase in Y as X_i increases by 1 unit, holding all other predictors fixed.

For our case, since we want to estimate the effect of advertising budgets of three different medias `TV`, `Newspaper`, and `Radio` onto `Sales`, then our multiple regression model can be represented like this

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

Estimating The Coefficients

Just like simple linear regression, we have to estimate the values of the regression coefficients $\beta_0, \beta_1, \beta_2, ..., \beta_n$, because they are unknown. Using `lm()` in R, we can directly get the estimated values $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 and \hat{\beta}_3$. Additionaly, we can also use the `summary()` function to summarize the regression object, to get more statistical details. (*results can be seen in Table 4*)

F-Statistic

Having the same concept with simple linear regression, it is important to know if there is actually a relationship between the response variable and the predictor variable. To determine wheter there is a relationship, one should conduct the null hypothesis to test if $\beta_0 = \beta_1 = \beta_2 = ... = \beta_n = 0$.

Null hypothesis :

$$H_0 : There is no relationship between X i and Y$$

$$H_0 : \beta_i = 0$$

versus

Alternative hypothesis:

$$H_a : There is some relationship between X i and Y$$

$$H_a : \beta_1 \neq 0$$

The hypothesis test is performed by computing the *F-statistic*,

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

where

$$TSS = \sum (y_i - \bar{y})^2$$

is the total sum of squares, and

$$RSS = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i 1 - \hat{\beta}_2 x_i 2 - ... - \hat{\beta}_n x_i n)^2$$

If the Null Hypothesis is true, then

$$E(TSS - RSS)/p = \sigma^2$$

, showing that when there is no relationship between the response and predictors the *F-statistc* take on a value close to 1. If the Alternative Hypothesis is true, then

$$E(TSS - RSS)/p > \sigma^2$$

, meaning that F is expected to be greater than 1 when there is some kind of relationship between the reponse and predictors.

Table 1: Summary TV

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.0326	0.4578	15.3603	0.0000
TV	0.0475	0.0027	17.6676	0.0000

Table 2: Summary of Multiple Linear Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9389	0.3119	9.4223	0.0000
TV	0.0458	0.0014	32.8086	0.0000
Radio	0.1885	0.0086	21.8935	0.0000
Newspaper	-0.0010	0.0059	-0.1767	0.8599

p-value

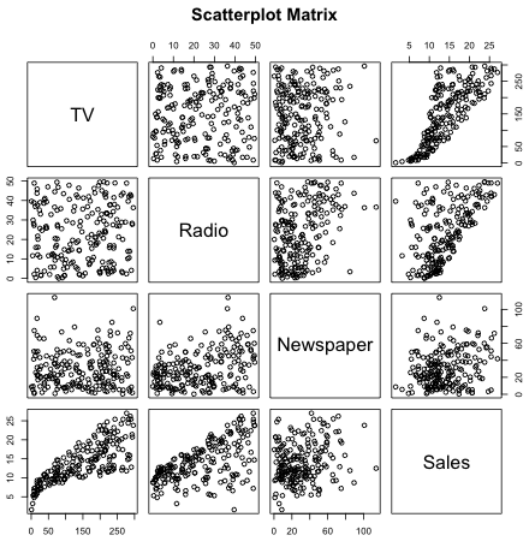
Results

1. Is at least one of the predictors useful in predicting the response? (regression of TV on sales) (regression of Radio on sales) (regression of Newspaper on sales)
2. Do all predictors help to explain the response, or is only a subset of the predictors useful?
3. How well does the model fit the data?

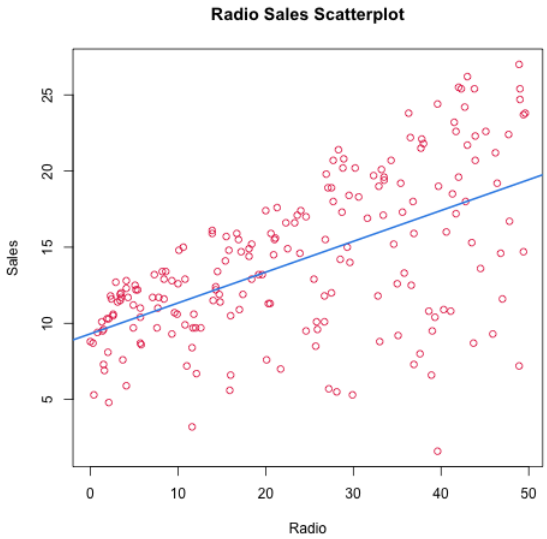
Table 3: Regression Quality Indices

	Quantity	Value
1	RSE	1.6855
2	R^2	0.8972
3	F-statistic	570.2707

last table 4. How accurate is the prediction?




TV Sales Scatterplot
Sales
TV
0
50
100
150
200
250
300
5
10
15
20
25


Radio Sales Scatterplot
Sales
Radio
0
10
20
30
40
50
5
10
15
20
25

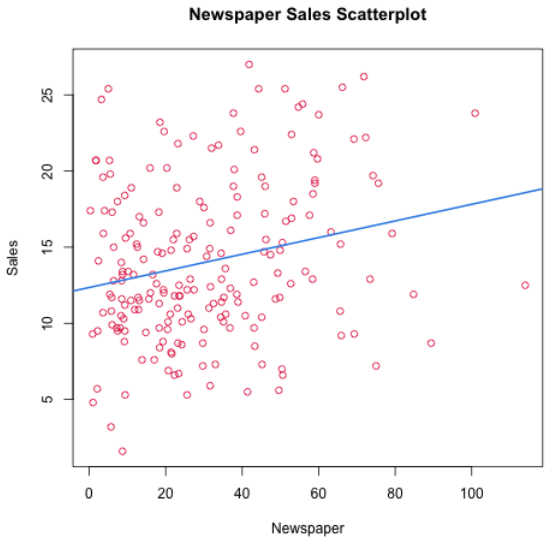


Table 4: Correlation Matrix for The 3 Medias

	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

```
print(xtable(TV_regression, caption="Regressing Sales on TV"))
```

```
## % latex table generated in R 3.2.2 by xtable 1.8-2 package
## % Fri Oct 14 05:19:01 2016
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\
##   \hline
## (Intercept) & 7.0326 & 0.4578 & 15.36 & 0.0000 \\
##   TV & 0.0475 & 0.0027 & 17.67 & 0.0000 \\
##    \hline
## \end{tabular}
## \caption{Regressing Sales on TV}
## \end{table}
```

Conclusions