Advertising Data Report

Lily Li

October 7, 2016

Abstract

This reports attempts to reproduce the results of advertising and sales found in section 3.1 Simple Linear Regression of *An Introduction to Statistical Learning*.

Introduction

The Advertising dataset contains data on sales (in thousands of units) for a particular product as a function of advertising budgets (in thousands of dollars) for TV, radio, and newspaper media. The goal is to suggest, on the basis of this data, a marketing plan for next year that will results in high product sales. In this report, we will focus how TV, radio, and newspaper budgets and their individual as well as combined relationship with sales. Some questions we would like to explore include:

1. Is at least one of the predictors useful in predicting the response?

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?

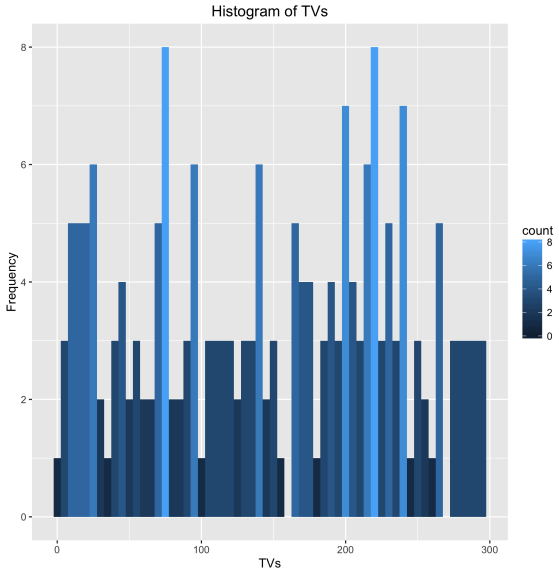
4. How accurate is the prediction?

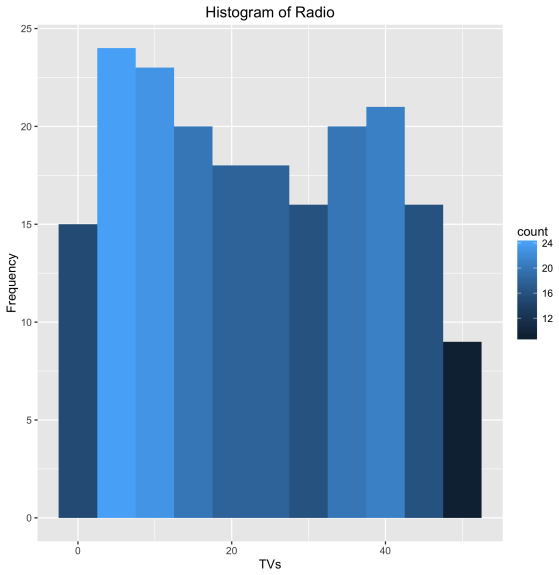
Data

This dataset has information on TV, Radio, and Newspaper budgets.
Some preliminary analysis of the dataset include:

-there are 200 observations of each TV, radio, newspaper, and sales
-histograms below show the distributions of TV, radio, newspaper, and sales data

Figure 1: Distributions of TV, Radio, Newspaper and Sales Data


Histogram of TVs
Frequency
TVs
count
0
2
4
6
8
0
100
200
300


Histogram of Radio
Frequency
TVs
count
0
5
10
15
20
25
0
20
40
12
16
20
24

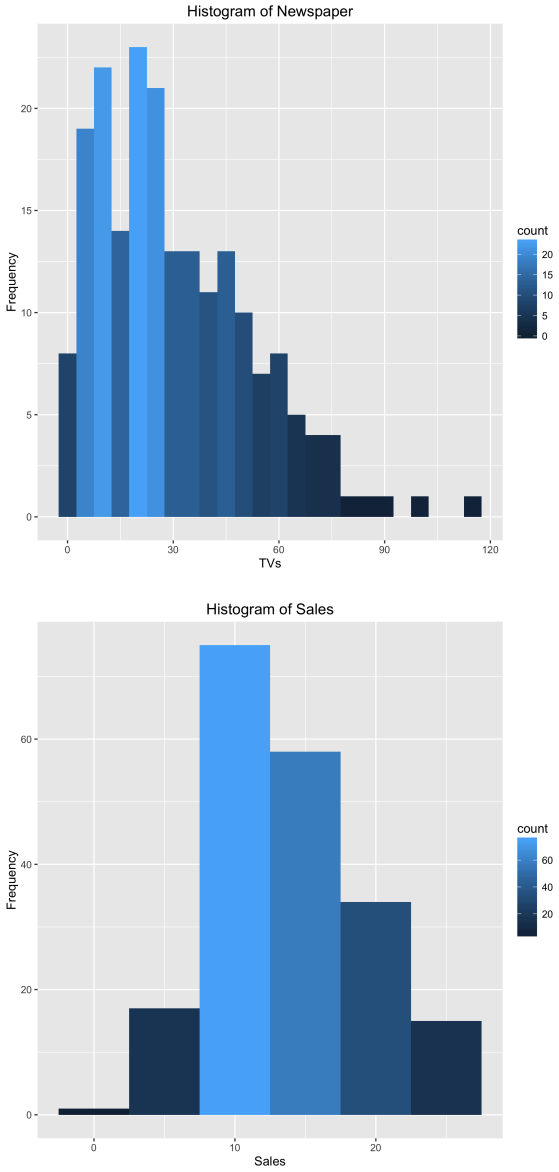


Figure 2: Correlation among the Variables

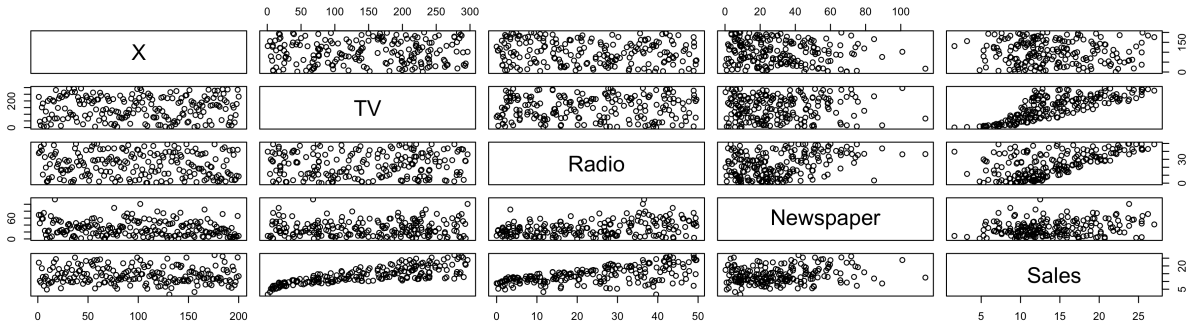


Figure 3: Correlation among the Variables Quantitative

	TV	Radio	Newspaper	Sales
TV	1	0.0548	0.0566	0.7822
Radio	0.0548	1	0.3541	0.5762
Newspaper	0.0566	0.3541	1	0.2283
Sales	0.7822	0.5762	0.2283	1

Methodology

We want to observe if there is a linear relationship between TV, Radio, and Newspaper budget and Sales. Let's consider the regression models:

$$Sales = \beta_{0,t} + \beta_{1,t} TV$$

$$Sales = \beta_{0,r} + \beta_{1,r} Radio$$

$$Sales = \beta_{0,n} + \beta_{1,n} Newspaper$$

$$Sales = \beta_{0,m} + \beta_{1,m} TV + \beta_{2,m} Radio + \beta_{3,m} Newspaper$$

To estimate the coefficients, we use the least squares minimization method.

Results

After computing the regression, we found the following results:

Table 1: TV Regression Coefficients

Table 2: Fitting linear model: Sales ~ TV

	Estimate	Std. Error	t value	Pr(>|t|)
TV	0.0475	0.0027	17.67	0
(Intercept)	7.033	0.4578	15.36	0

Figure 1: Scatterplot of TV Regression

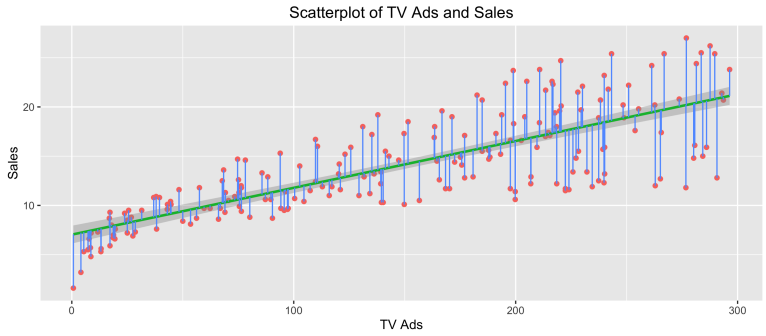


Table 2: Radio Regression Coefficients

Table 3: Fitting linear model: Sales ~ Radio

	Estimate	Std. Error	t value	Pr(>|t|)
Radio	0.2025	0.0204	9.921	0
(Intercept)	9.312	0.5629	16.54	0

Figure 2: Scatterplot of Radio Regression

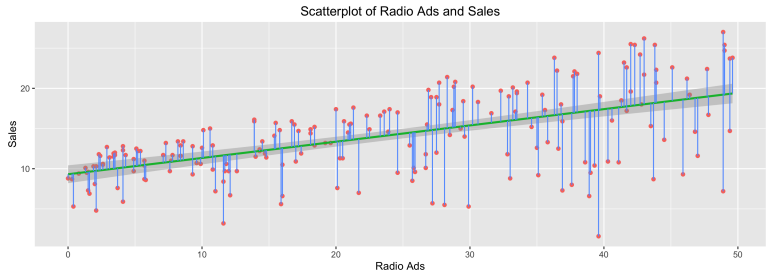


Table 3: Newspaper Regression Coefficients

Table 4: Fitting linear model: Sales ~ Newspaper

	Estimate	Std. Error	t value	Pr(>|t|)
Newspaper	0.0547	0.0166	3.3	0.0011
(Intercept)	12.35	0.6214	19.88	0

Figure 3: Scatterplot of Newspaper Regression

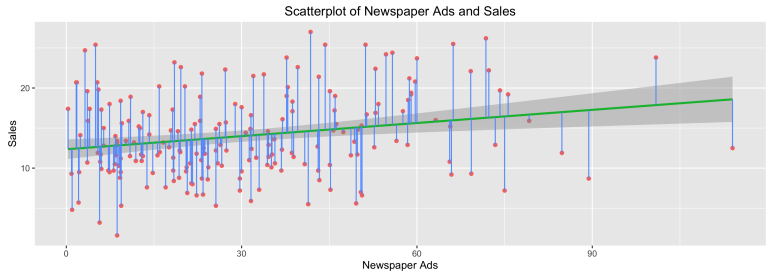


Table 4: Multiple Regression Coefficients

Table 5: Fitting linear model: Sales ~ TV + Radio + Newspaper

	Estimate	Std. Error	t value	Pr(>|t|)
TV	0.0458	0.0014	32.81	0
Radio	0.1885	0.0086	21.89	0
Newspaper	-0.001	0.0059	-0.1767	0.8599
(Intercept)	2.939	0.3119	9.422	0

Table 5: Regression Quality Statistics

	RSS	R Square	F Stat
TV	2103	0.6119	312.1
Radio	3618	0.332	98.42
Newspaper	5135	0.0521	10.89
Multiple	556.8	0.8972	570.3

Conclusions

Looking at the *TV* regression coefficent in *Table 1* shows a positive relationship with *sales*. In fact, 1% increase in *TV* budget is associated with 4.75% increase in *sales*. The t statistic of this coefficient is so large that the p-value is less than 4 decimal places of 0. This coefficient is statistically significant.

The R squared in *Table 5* shows that 61% of the variation in sales is explained by the variation in TV budget.

1. Is at least one of the predictors useful in predicting the response?
   Yes, when running single regressor regressions, each regressor were found to be significant at the 5% level.
2. Do all predictors help to explain the response, or is only a subset of the predictors useful?
   Individually, the regressors seem significant. However, when running a multiple regression, which has the highest F stat, the Newspaper regressor seems to have a negative relationship with sales. But this finding should be judged lightly because there's no statistical significance with a pvalue of 0.8599151.
3. How well does the model fit the data?
   The multiple regression model has an r squared of 0.8972106 meaning 89.7210638 of the variation in sales is explained by the variation in TV, radio, and newspaper budget. Compared to the individual models of TV, radio, and newspaper, which has r squared of 0.6118751, 0.3320325, and 0.0521204 (respectively), the multiple regression model has the best variation explaination.
4. How accurate is the prediction?
   Accuracy is difficult to say with the limitations of this specific report. Cross validation and accuracy proportions would be explored.