

Report

Atul Lanka

October 14th, 2016

Abstract

This report will replicate the main results displayed in section 3.2 **Multiple Linear Regression** (chapter 3) of the book *An Introduction to Statistical Learning*.

Introduction

The primary goal of this analysis is to give advice on how to boost sales of the product given the current information on advertising budgets. More specifically, the idea is to determine whether there exists a correlation between advertising and sales, and if so, formulate an accurate model that can be used to predict sales from media budget. For this analysis in particular, a combination of simple linear regression and multiple linear regression.

Data

The Advertising data set comprises of the Sales (in thousands of units) in 200 different markets, along with the advertising budgets (in thousands of dollars) in each market for three different forms of media: TV, Radio, and Newspaper. In this report, the relation between each of them and Sales, and the possible relation between Sales and the three of them combined are observed and studied/analyzed.

We may first look at the table of summary statistics below:

	TV	Radio	Newspaper	Sales
1	Min. : 0.70	Min. : 0.000	Min. : 0.30	Min. : 1.60
2	1st Qu.: 74.38	1st Qu.: 9.975	1st Qu.: 12.75	1st Qu.:10.38
3	Median :149.75	Median :22.900	Median : 25.75	Median :12.90
4	Mean :147.04	Mean :23.264	Mean : 30.55	Mean :14.02
5	3rd Qu.:218.82	3rd Qu.:36.525	3rd Qu.: 45.10	3rd Qu.:17.40
6	Max. :296.40	Max. :49.600	Max. :114.00	Max. :27.00

Table 1: Summary Statistics

Histograms for each variable:

Methodology

Single Linear Regression

We consider each media separately from the data set - TV, Radio and Newspaper - and study its relationship with the dependent variable Sales. The null hypothesis is that each of the independent variables would not have an effect on Sales, and the alternate hypothesis suggests otherwise. Thus a linear model is generated:

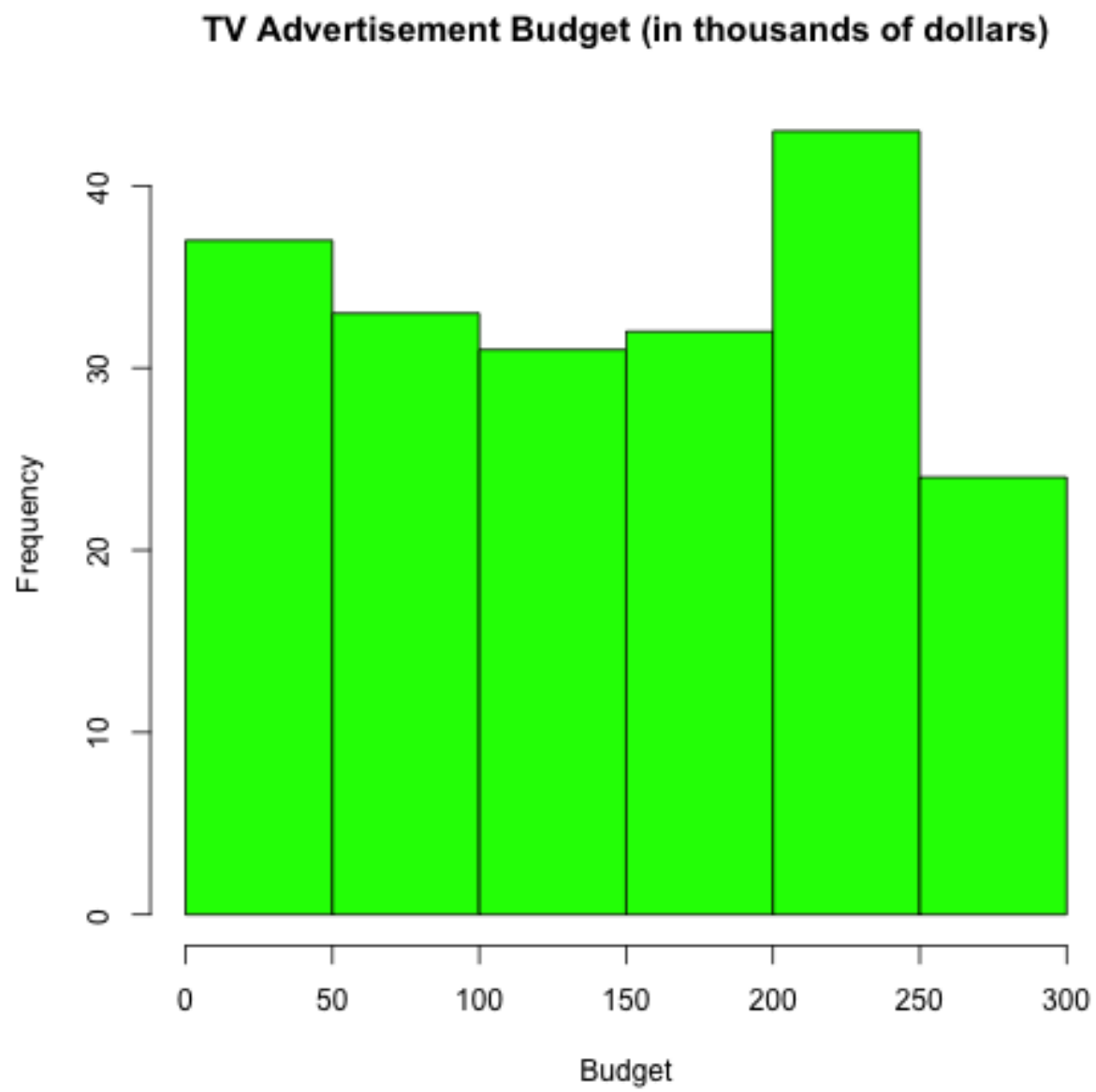


Figure 1: Figure 1: Histogram for TV

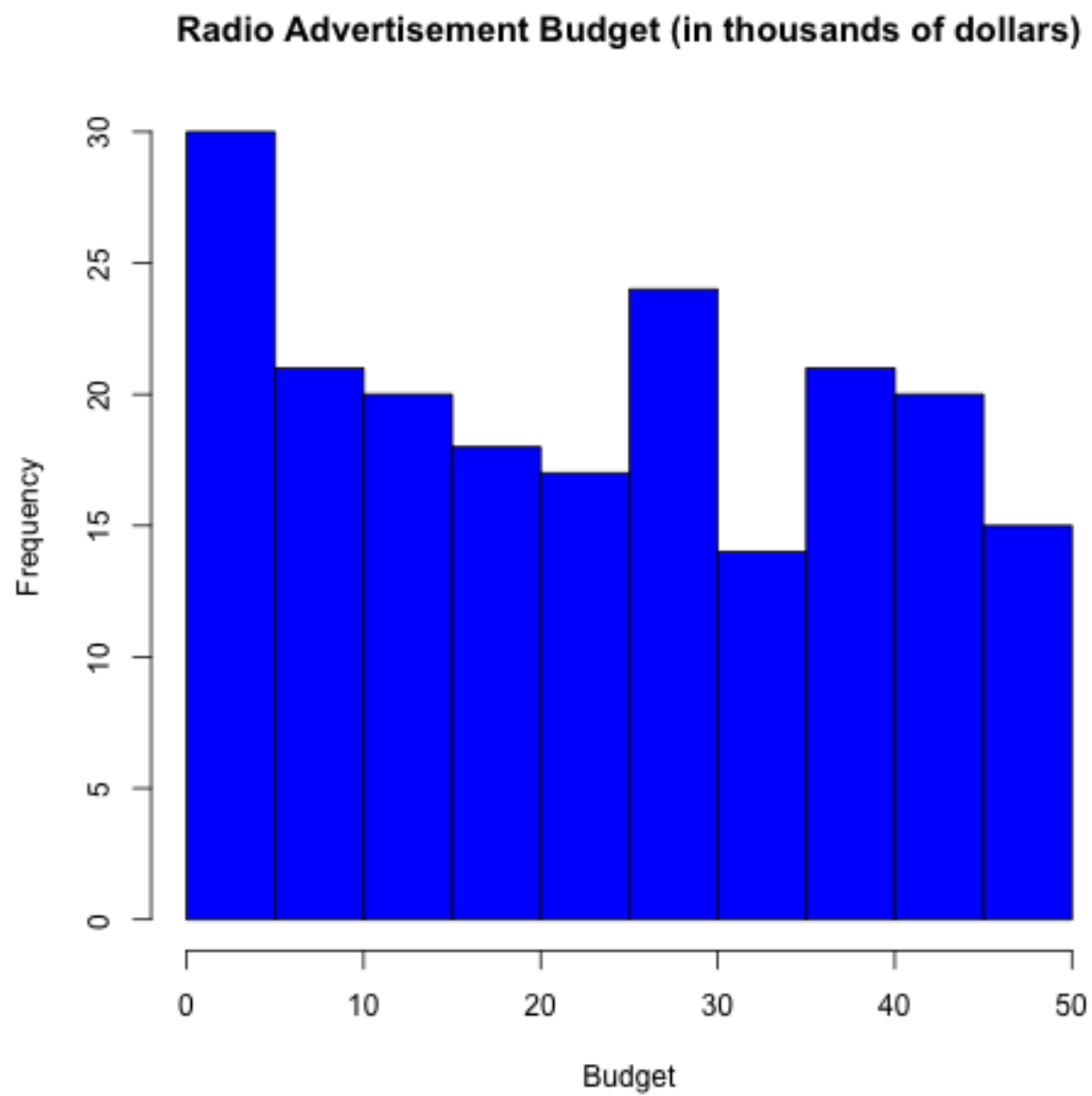


Figure 2: Figure 2: Histogram for Radio

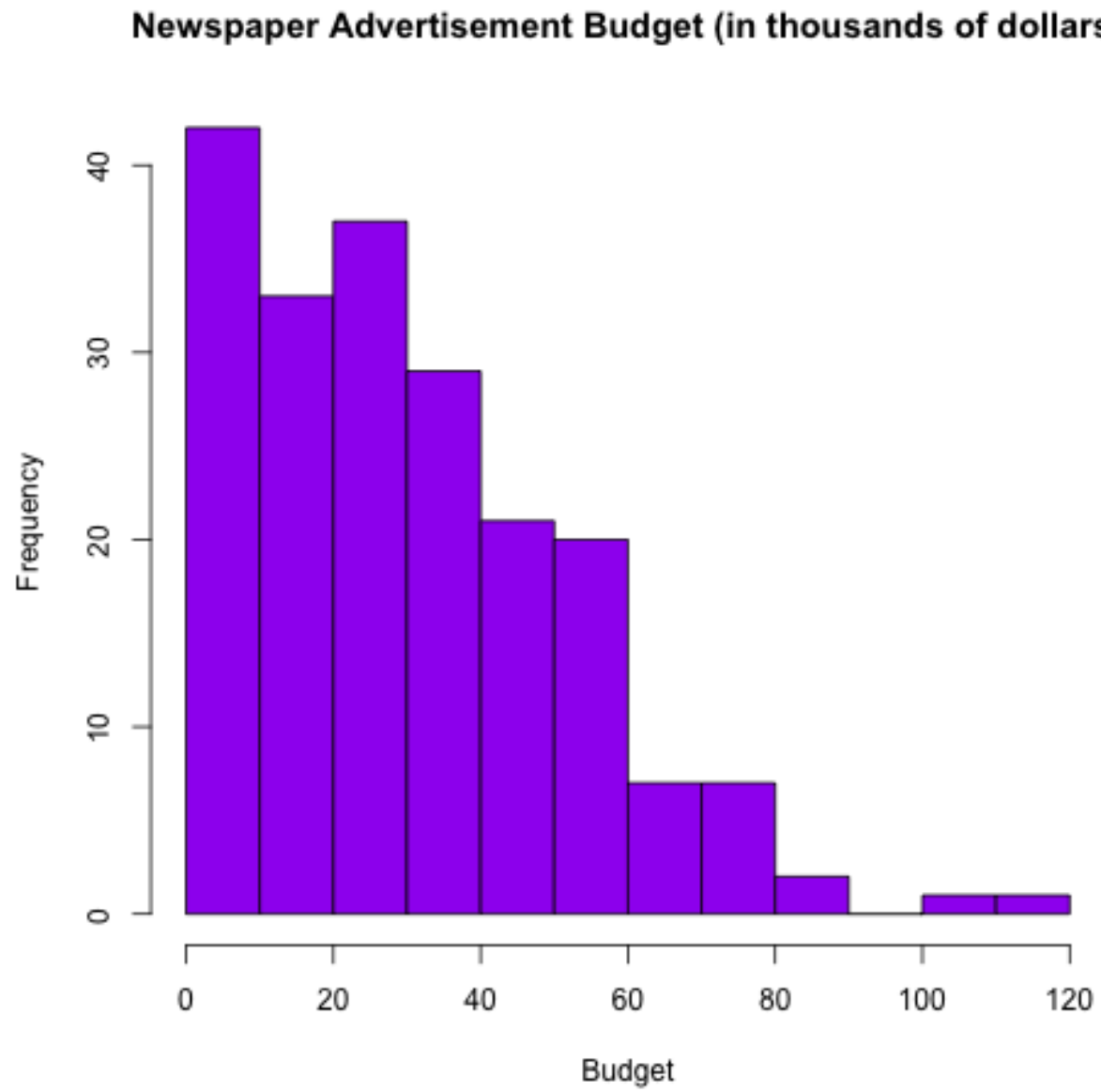


Figure 3: Figure 3: Histogram for Newspaper

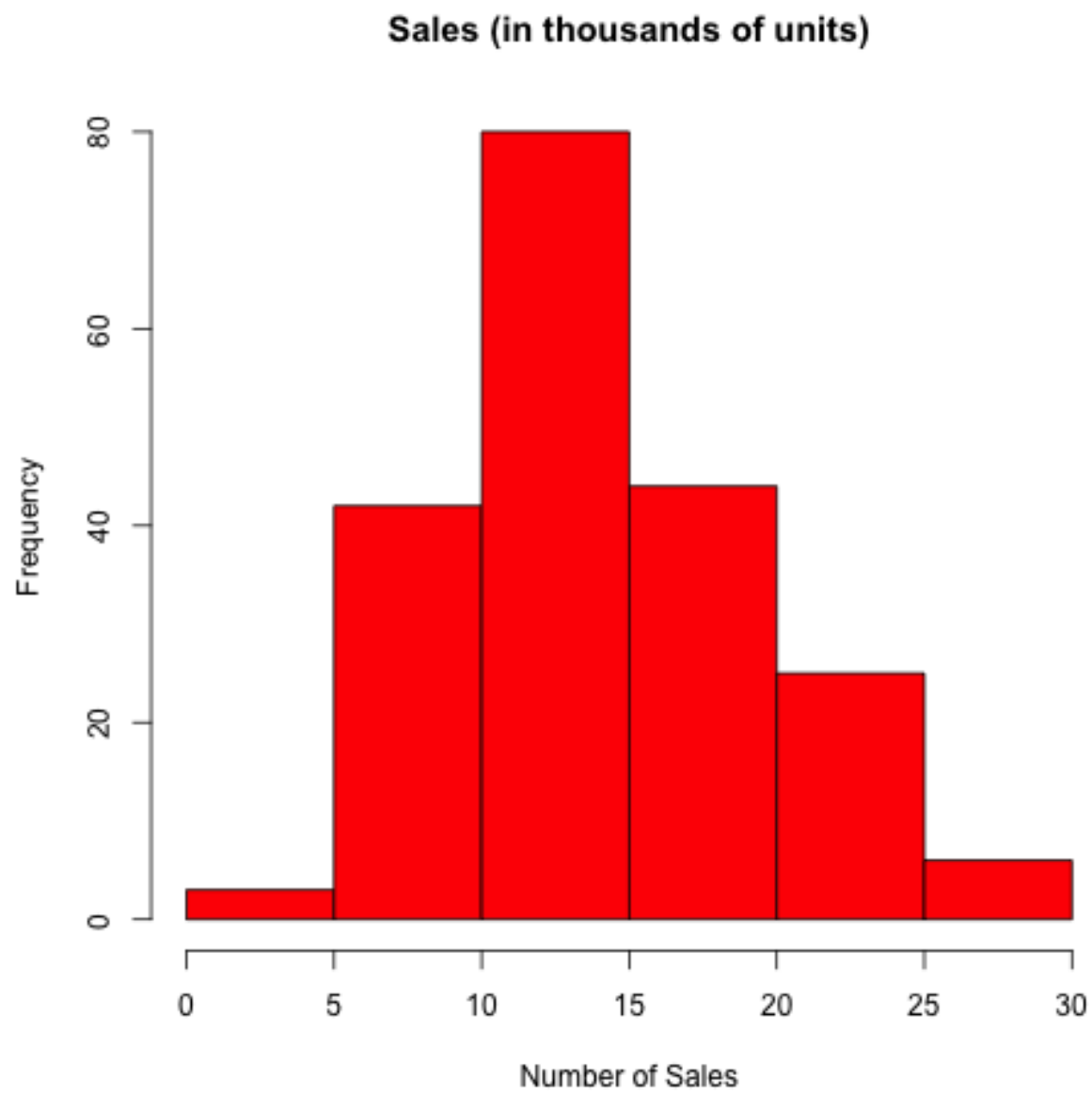


Figure 4: Figure 4: Histogram for Sales

$$Sales = \beta_0 + \beta_1(TV|Radio|Newspaper)$$

Multiple Linear Regression

It would be a better approach to expand the model of Sales with multiple predictors rather than single predictors each time, avoiding an excess of linear models. This is done by accomodating each predictor with their respective slope coefficient in a single model. Thus the multiple linear regression model takes the form:

$$Sales = \beta_0 + \beta_1TV + \beta_2Radio + \beta_3Newspaper$$

Results

Single Linear Regression

Using `lm()` to fit the data into a simple linear model, the regression coefficients are as follows:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.0326	0.4578	15.36	0.0000
advert\$TV	0.0475	0.0027	17.67	0.0000

Table 2: Regression Coefficients for TV Sales

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.3116	0.5629	16.54	0.0000
advert\$Radio	0.2025	0.0204	9.92	0.0000

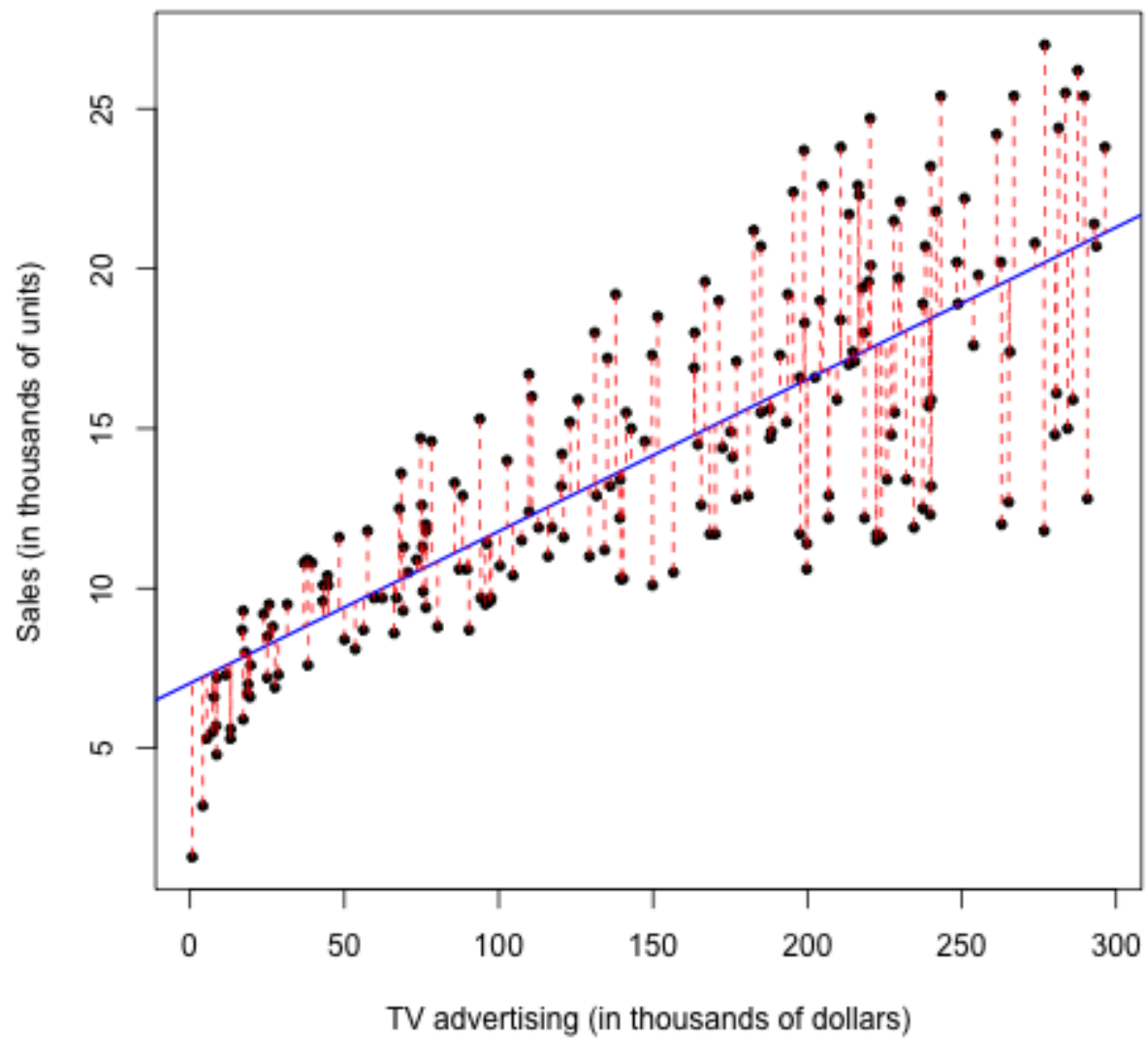
Table 3: Regression Coefficients for TV Radio

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	12.3514	0.6214	19.88	0.0000
advert\$Newspaper	0.0547	0.0166	3.30	0.0011

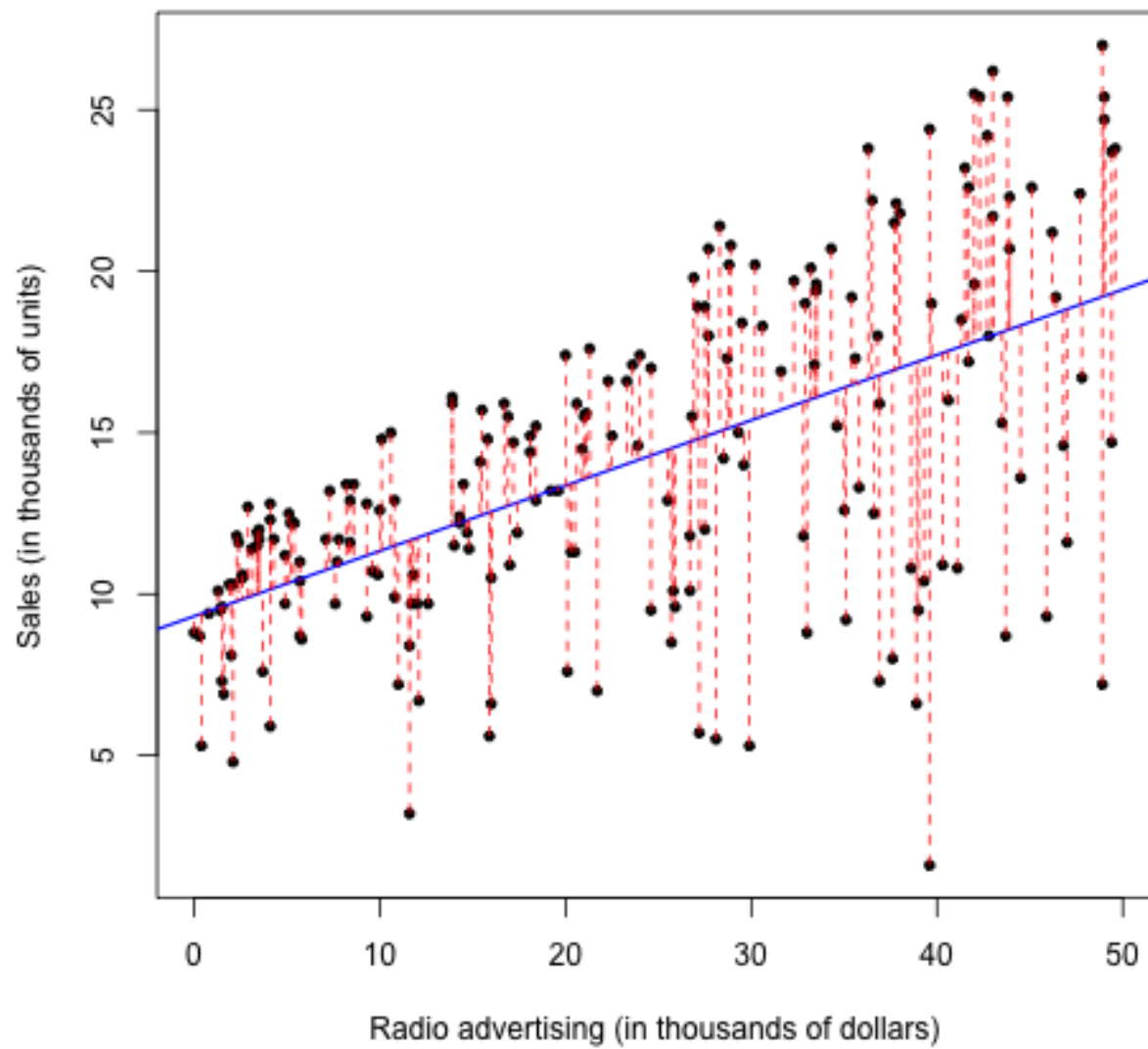
Table 4: Regression Coefficients for TV Newspaper

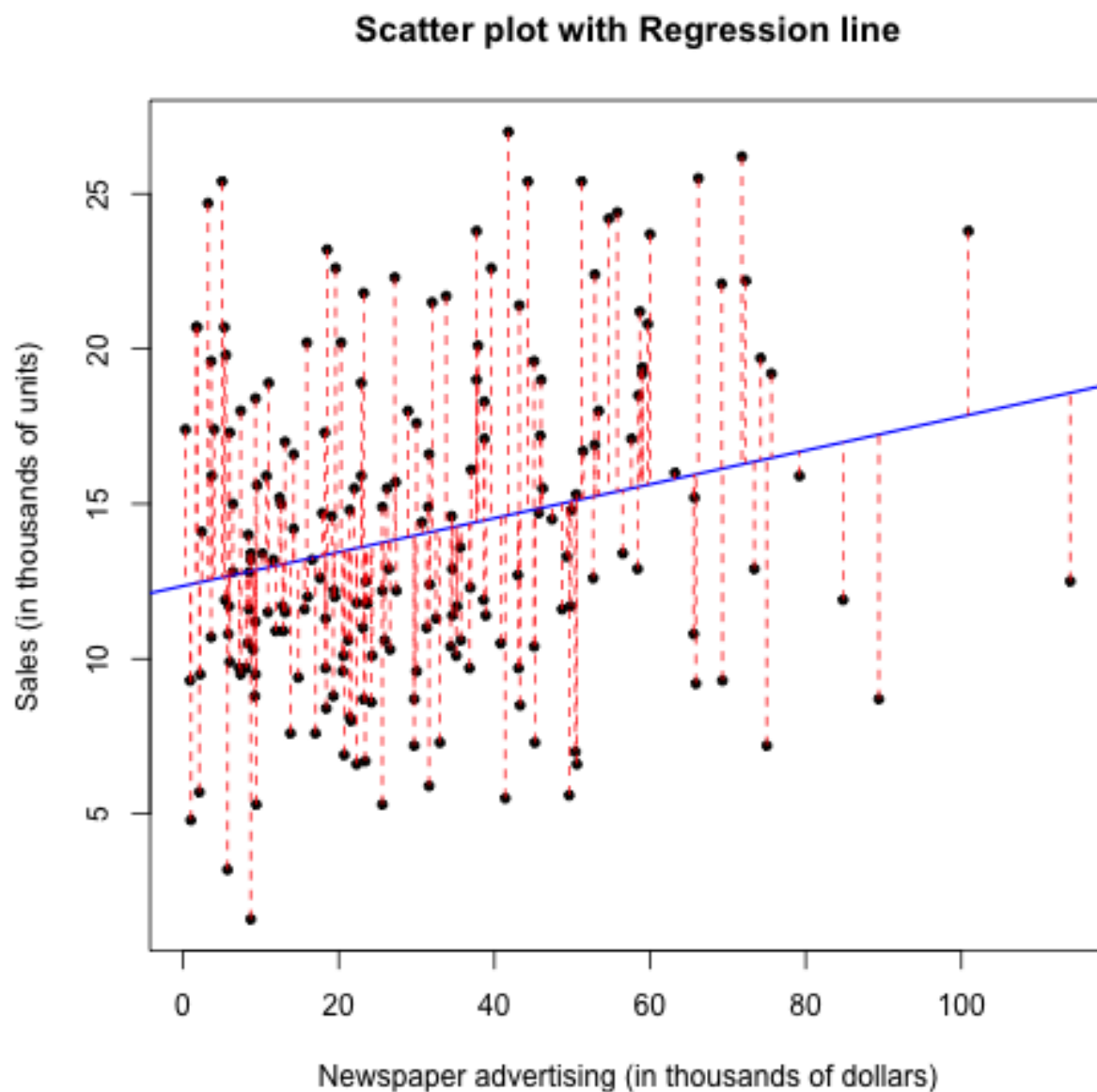
The scatterplots with their respective regression lines are as follows:

Scatter plot with Regression line



Scatter plot with Regression line





Quality Indices calculated with the regression analysis are as follows:

	Quantity	Value
1	RSE	3.25865636865046
2	R2	0.611875050850071
3	F-stat	312.144994372713

Table 5: Quality Indices of Regression of Sales on TV

Conclusions

In conclusion, the multiple linear regression is more accurate than the single linear regressions. From the tables above (**F-statistic**) from multiple regression, at the very least one of the predictors can be used

to predict **Sales**. Nevertheless, we also find that not all of the predictors are statistically significant (from the p-values). Hence, the prediction would be more accurate if the **newspaper** budget is not avoided based on its corresponding p-value. Other indicators including Residual Standard Error, R^2 and F-statistic also comment on the fit of the linear model - the smaller these values/statistics or approaching 0, the more accurately the model represents the data.