Regression Analysis on the Relationship Between Advertising Budgets and Product Sales

Shannon Chang

October 14, 2016

Abstract

In this report, I will reproduce the multiple linear regression analysis detailed in Chapter 3 of Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani's *An Introduction to Statistical Learning*. The analysis is carried out on the Advertising.csv dataset that is paired with the textbook, and contains data on product sales in over two hundred different markets along with the advertising budgets for the product in each market by different mediums: TV, Radio, and Newspaper.

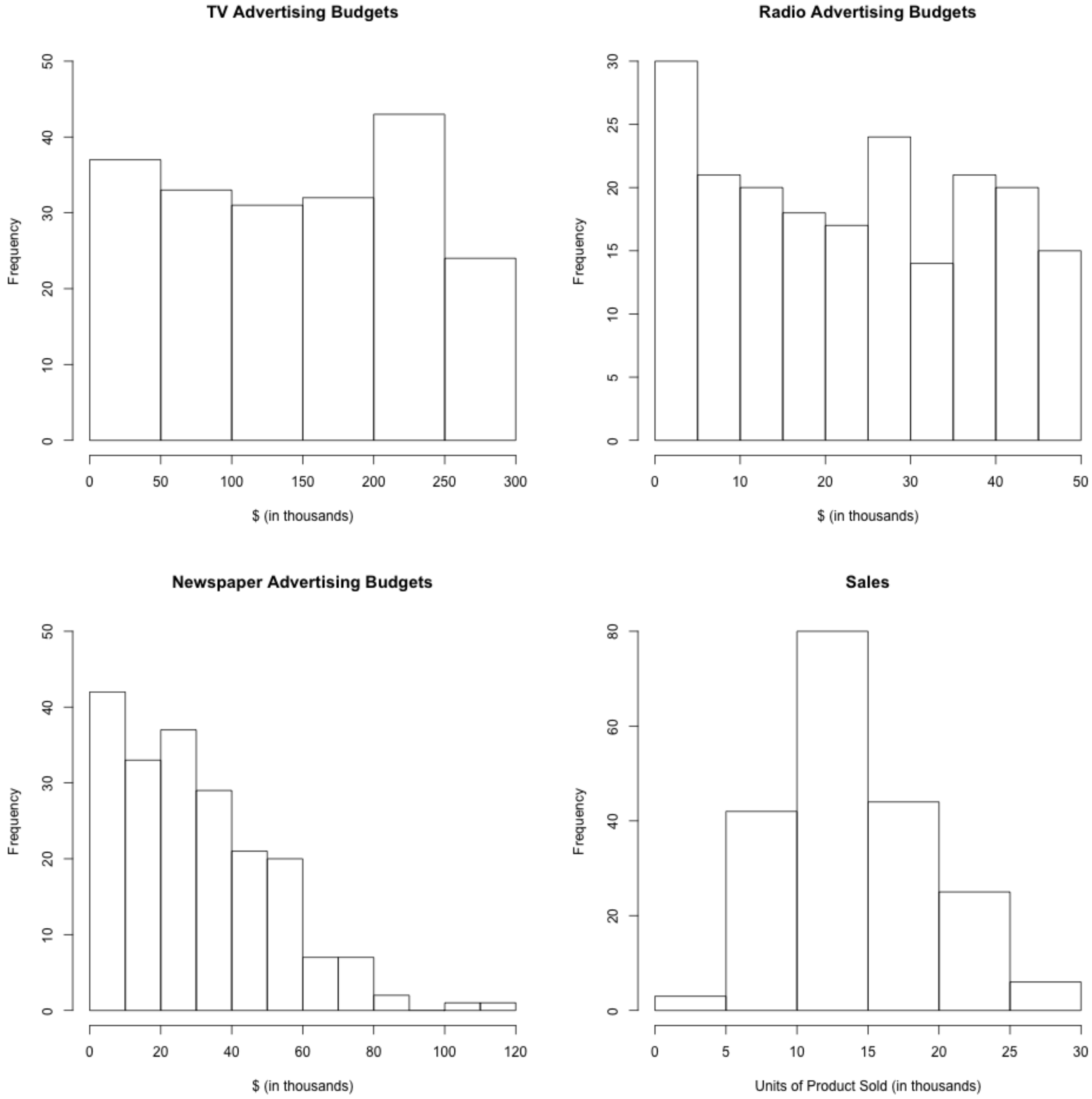
Specifically, I will reproduce:
* **Table 3.1** (page 72) - Coefficient estimates of simple linear regression models of Sales on TV, Sales on Radio, and Sales on Newspaper
* **Table 3.4** (page 74) - Least squares coefficient estimates of the multiple linear regression model of Sales on TV, Radio, and Newspaper
* **Table 3.5** (page 75) - A correlation matrix for TV, Radio, Newspaper, and Sales
* **Table 3.6** (page 76) - RSE, R^2, and $F-statistic$ values from the least squares model for the regression of Sales on TV, Radio, and Newspaper

Introduction

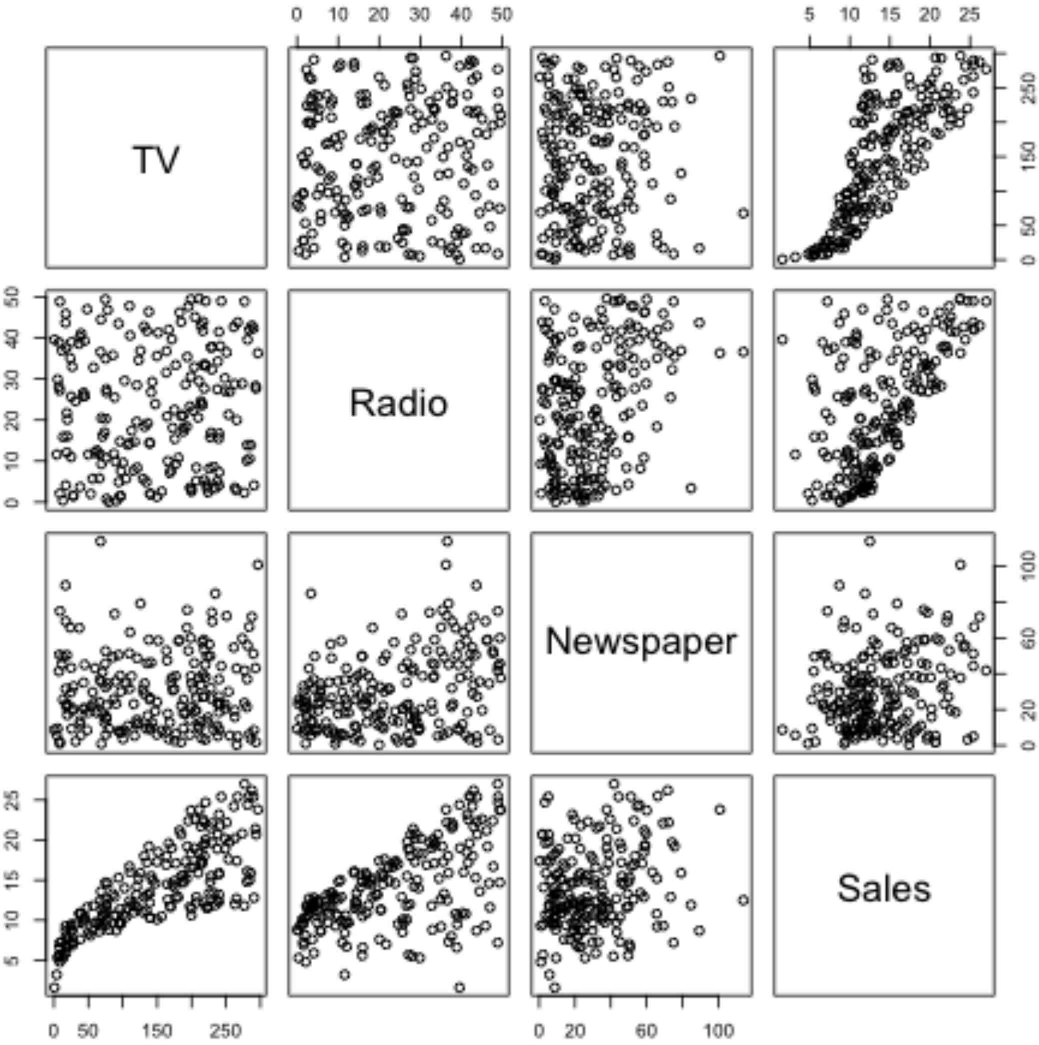
Suppose a company wants advice on how to increase sales for one of its products. There is, of course, no concrete way to insure increased sales, but we can influence greater sales through advertising. Imagine that we are statistical consultants hired for this project. To convince the company to invest in advertising campaigns, we must first prove to the company that there is a relationship between advertising budget and sales. From here, we can then advise the company on appropriate budgets to better reach sales targets. Thus, the goal for this analysis is to determine whether there is a relationship between advertising budget and sales and, if so, how strong the relationship is between the two. Given that there is a relationship between advertising budget and sales, we would want to construct an accurate model that can be utilized to predict sales based on the size of advertising budget. There are three mediums detailed in Advertising.csv: TV, Radio, and Newspaper, so we will examine their individual relationships with Sales and then their combined relationship with Sales.

Data

The Advertising.csv dataset consists of advertising budgets (in thousands of dollars) by medium: TV, Radio, and Newspaper. Product sales (in thousands of units) are listed under Sales. There are 200 rows of data, indicating 200 different markets. A very general overview of the distribution for each column of data can be seen in the histograms below:
Histograms for the two columns are shown below:



A scatterplot matrix for the entire `Advertising.csv` dataset is shown below:



Methodology

Setting Up a Model

To examine the association between `Sales` and `TV`, `Radio`, and `Newspaper`, whether individually or altogether, we model the relationship by *linear regression.* For the association with `Sales` and the advertising budget for each medium, we would model the relationship by *simple linear regression.* For the association between `Sales` and all three mediums, we would model the relationship by *multiple linear regression.*

Simple Linear Regression

This method involves predicting a quantitative response Y, based on the predictor variable X. For the association between `Sales` and the advertising budget for each medium, we would model the relationships as:

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

$$Sales = \beta_0 + \beta_1 Radio$$

$$Sales = \beta_0 + \beta_1 Newspaper$$

Here, Here, β_0 represents the intercept of the linear model while β_1 represents the slope.

Since all β_0 and β_1 are unknown for each association, we would need to calculate estimates for the their coefficients instead. In a visual sense, we would want to graph all the data for `TV` and `Sales` and fit a line $Sales = \beta_0 + \beta_1 TV$ as close as possible to our 200 data points. We would repeat this process for `Radio` and `Newspaper` as well. Then, we can optimize the fit of each line using the *least squares criterion.* This involves minimizing the *residual sum of squares*; a *residual* is the distance between each data point and its predicted value from the linear model). The linear model/line that we fit would be based on an average of the squares.

From a computational perspective, we can start fitting the line by calculating estimates $\hat{\beta_0}$ and $\hat{\beta_1}$ for β_0 and β_1. $\hat{\beta_0}$ and $\hat{\beta_1}$ are chosen to minimize the aforementioned *residual sum of squares* (RSS), which is given by the equation:

$$RSS = (y_1 - \hat{\beta_0} - \hat{\beta_1} x_1)^2 + (y_2 - \hat{\beta_0} - \hat{\beta_1} x_2)^2 + ... + (y_n - \hat{\beta_0} - \hat{\beta_1} x_n)^2$$

$\hat{\beta_0}$ and $\hat{\beta_1}$, in turn, are shown through caculus derivation as:

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$$

$$\hat{\beta_1} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Here, $\bar{x}$ is the mean of `TV` (or `Radio` or `Newspaper`), while $\bar{y}$ is the mean of `Sales`.

To evaluate the accuracy of these estimates, we start by calculating standard errors of the standard means. We can then use these standard errors to perform *hypothesis tests* on the estimates. Thus, we would be testing the *null hypothesis* that

$$H_0 : There\, is\, no\, relationship\, between\, TV\, and\, Sales$$

versus the *alternative hypothesis* that

$$H_1 : There\, is\, some\, relationship\, between\, TV\, and\, Sales$$

Here, TV can be substituted with *Radio* and *Newspaper*.

Numerically, we would be testing

$$H_0 : \beta_1 = 0$$

versus

$$H_0 : \beta_1 \neq 0$$

To do so, we would calculate a *t-statistic*:

$$t = \frac{\hat{\beta_1} - 0}{SE(\hat{\beta_1})}$$

This measures the number of standard deviations that our estimate for β_1 is away from 0.

From this, we can calculate a *p-value*, which is the probability of observing any value greater than or equal to $|t|$. A small p-value would indicate that it is unlikely to observe a meaningful association between the predictor (TV, or Radio, or Newspaper) and the response (Sales) purely by chance without some true relationship between the two. Thus, a small p-value would allow us to *reject the null hypothesis* and determine that there is a relationship between TV (or Radio or Newspaper) and Sales. In general, 5% or 1% are used as p-value benchmarks.

Muliple Linear Regression

This method is, in many ways, similar to **simple linear regression**, but with the added complexity of more variables. We start by predicting a quantitative response Y, based on the predictor variables X_1, X_2, and X_3. For the combined association between Sales and TV, Radio, and Newspaper, we would model this relationship as:

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

Here, Here, β_0 represents the intercept of the linear model while β_1, β_2, and β_3 represent the three different slopes. Each β_j can be interpreted as *the average effect on Y of a one unit increase in X_j, holding all other predictors fixed.*Note that fitting three simple linear regressions is not the same as fitting one multiple linear regression for Sales, because the each simple linear regression would not take into account the effects of the other two mediums and, subsequently, any interaction effects.

Visualizing this model is a complicated matter, as each additional predictor adds another dimension to the graph. But much like in the *simple linear regression*, we are once again trying the minimize the sum of squared residuals, this this time with an expanded formulation:

$$\begin{aligned} RSS &= \sum_{i=1}^{n}(y_i - \hat{y_i})^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{\beta_0} - \hat{\beta_1}x_{i1} - \hat{\beta_p}x_{i3}) \end{aligned}$$

The coefficient estimates of $\hat{\beta_0}$, $\hat{\beta_1}$, $\hat{\beta_2}$, and $\hat{\beta_3}$ are derived more complicatedly through matrix algebra and will not be listed here.

Hypothesis testing for similar to that conducted for *simple linear regression* can also be conducted by *multiple linear regression.* The calculated results, however, will change because of the way in which $\hat{\beta_0}$, $\hat{\beta_1}$, $\hat{\beta_2}$, and $\hat{\beta_3}$ are calculated. Hence, results of a hypothesis test conducted from the *multiple linear regression*, say for Sales and TV could have different results compared to a hypothesis conducted from a *simple linear regression* for the same variables.

In general for *multiple linear regression*, however, we want to test whether all of the regression coefficients are zero, i.e. whether $\beta_1 = \beta_2 = \beta_3 = 0$ in this case. Hence, we would test the null hypothesis that

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

versus the *alternative hypothesis* that

$$H_1 : At\,least\,one\,\beta_j\,is\,non-zero.$$

This hypothesis would be conducted by calculating the $F-statistic$:

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

(TSS is further explained in the sections below.)

We can also test the hypothesis that a certain subset q of the coefficients are 0. For this, we would test the null hypothesis that

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \beta_p = 0$$

where p is the number of predictors, which in this case is 3. The $F - statistic$ would then be calculated by:

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

where RSS_0 is the residual sum of squares for a second model that uses all the variables _except_those last q.

Evaluating Accuracy of the Model

After conducting a hypothesis test, we will want to examine the extent to which the model fits the data. There are three values that we can look at to assess this: the *residual standard error* (RSE), the R^2 value, and the $F - statistic$.

Residual standard error (RSE)

The RSE is an estimate of the standard deviation of errors, the distances from each data point to its predicted value based on the linear model we fit. In other words, it is the average amount that the response (`Sales`) will differ from the true regression line and is given by the formula:

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum (y_i - \hat{y_i})^2}$$

R^2 value

The R^2 statistic is, technically speaking, the proportion of variance explained by our fitted model. Specifically, it measures the *proportion of variability in the response (`Sales`) that can be explained using the predictor (`TV`)*. The closer R^2 is to 1, the greater the proportion of variability that is explained. Its formula is given by:

$$R^2 = \frac{(TSS - RSS)}{TSS} = 1 - \frac{RSS}{TSS}$$

Here, the *total sum of squares*, TSS $= \sum (y_i - \bar{y})^2)$ measures the total variance in the response Y, and can be thought of as the amount of variability that already exists in the response, even before we perform any regression analysis. Thus, the Rˆ2 value is a ratio of variability in Y that can be explained by our model to the variability that exists inherently in Y.

F-statistic

Following up on the earlier discussion of the $F - statistic$, it can be shown by linear model assumptions that

$$ERSS/(n-p-1) = \sigma^2$$

Provided that H_0 is true,

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

Thus, when there is no relationship between the response and its predictors, we can expect the $F - statistic$ to be close to or equal to 1. If H_1 is true, on the other hand, then $E(TSS - RSS)/p > \sigma^2$, and we can expect F to be greater than 1.

Results

Simple Linear Regression

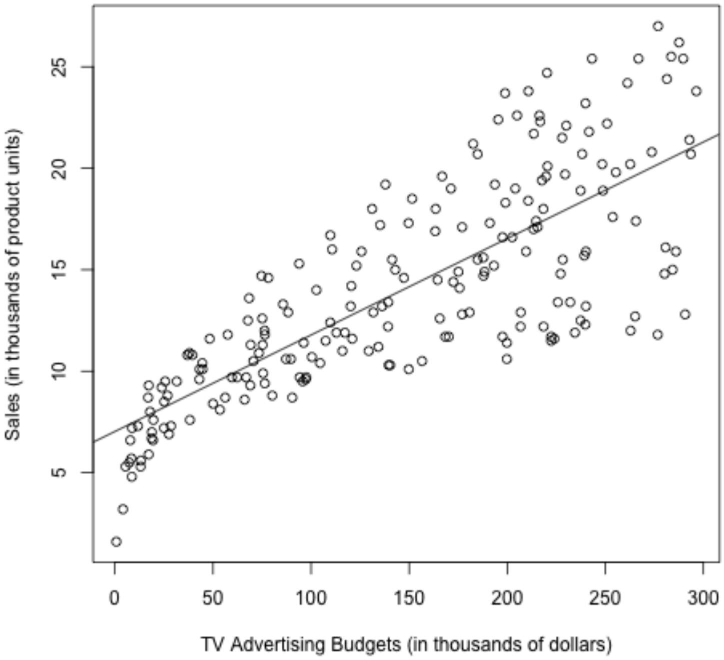
For the association between `Sales` and `TV`, I used R to calculate a regression object, which produced the coefficient estimates for the model and necessary calculations for performing a *t-test*. I also plotted the observed data against the line fitted by the regression object in order to obtain a visualization of this analysis. The code is located in ‘code/scripts/regression-script.R" of the repository for this paper, and is summarized by the following:

```r
# Load necessary package(s) and data
library(readr)
advertising <- read.csv(file = "../data/Advertising.csv", row.names = 1)

# Generate regression object
sales_tv_reg <- lm(Sales ~ TV, data = advertising)

# Generate summary
sales_tv_sum <- summary(sales_tv_reg)

# Generate plot
plot(x = advertising$TV, y = advertising$Sales,
     xlab = "TV Advertising Budgets (in thousands of dollars)",
     ylab = "Sales (in thousands of product units)")
abline(sales_tv_reg)
```

This generated the following outputs:

Table 1: Coefficient Estimates from the Simple Regression of Sales on TV

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.033	0.458	15.36	0.0000
TV	0.048	0.003	17.67	0.0000

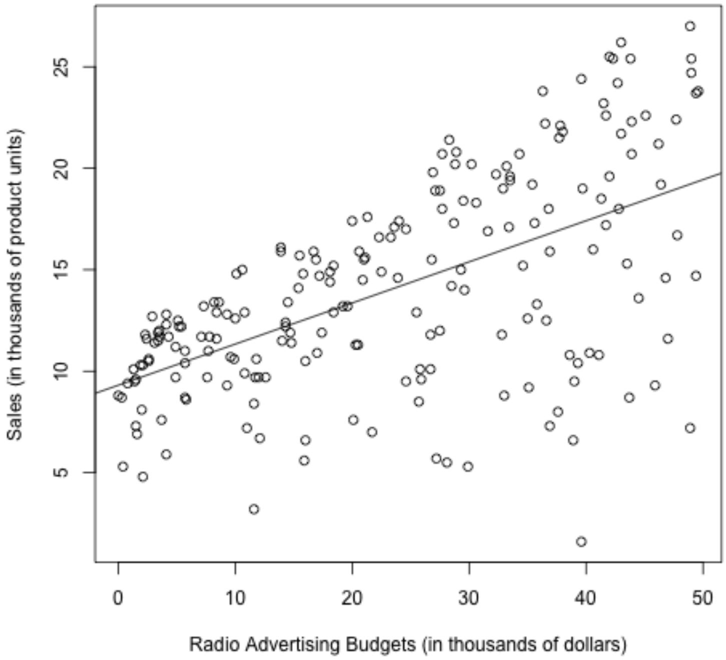


From **Table 1**, we can see that the coefficient estimate for β_0 is 7.033, while β_1 is 0.048. The p-value associated with `TV` is 0, which is less than .01 and .05–meaning that the coefficient estimate for `TV` is statistically significant at the 1% and 5% level. This means that we can reject the null hypothesis that there is no relationship between `Sales` and `TV`.

Using the coding process detailed above, I was further able to produce analysis outputs for the association between `Sales` and `Radio` and `Sales` and `Newspaper`:

Table 2: Coefficient Estimates from the Simple Regression of Sales on Radio

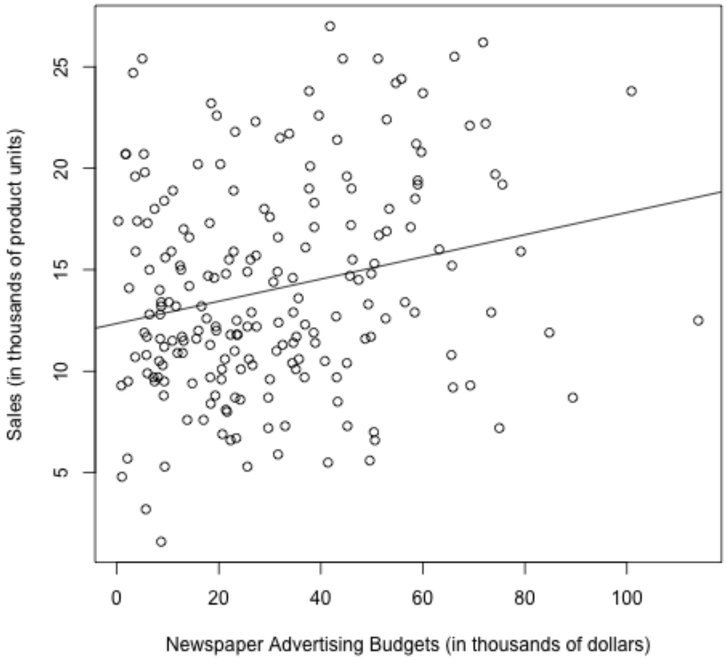
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.312	0.563	16.54	0.0000
Radio	0.202	0.020	9.92	0.0000



From **Table 2**, we can see that the coefficient estimate for β_0 is 9.312, while β_1 is 0.202. The p-value associated with `Radio` is 0, which is less than .01 and .05–meaning that the coefficient estimate for `TV` is statistically significant at the 1% and 5% level. This means that we can reject the null hypothesis that there is no relationship between `Sales` and `Radio`.

Table 3: Coefficient Estimates from the Simple Regression of Sales on Newspaper

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.351	0.621	19.88	0.0000
Newspaper	0.055	0.017	3.30	0.0011



From **Table 3**, we can see that the coefficient estimate for β_0 is 12.351, while β_1 is 0.055. The p-value associated with `Newspaper` is 0.0011, which is less than .01 and .05–meaning that the coefficient estimate for `TV` is statistically significant at the 1% and 5% level. This means that we can reject the null hypothesis that there is no relationship between `Sales` and `Newspaper`.

Multiple Linear Regression

For the combined association between `Sales` and `TV` and `Radio` and `Newspaper`, I first used R to calculate the a correlation matrix to get a very general overview of the relationship betwen all four variables. (Keep in mind, however, the old saying that "correlation does not imply causation.") I then calculated a regression object, which produced the coefficient estimates for the model and necessary calculations for performing a *t-test*. I also generated diagnostic plots of the regression, which a plot of the `Sales` residuals against fitted values, a Scale-Location plot of squared residuals against fitted values, and a normal Q-Q plot. The code is located in 'code/scripts/regression-script.R" of the repository for this paper, and is summarized by the following:

```
# Load necessary package(s) and data
library(readr)
advertising <- read.csv(file = "../data/Advertising.csv", row.names = 1)

# Generate correlation matrix
adv_cor <- cor(advertising, use = "all.obs")

# Generate regression object
multi_reg <- lm(Sales ~ TV + Radio + Newspaper, data = advertising)

# Generate summary
multi_sum <- summary(multi_reg)
```

```
# Generate plots
# Residual plot
plot(multi_reg, which = c(1))

# Scale location plot
plot(multi_reg, which = c(3))

# Normal qq plot
plot(multi_reg, which = c(2))
```

This generated the following outputs:

Table 4: Correlation Matrix

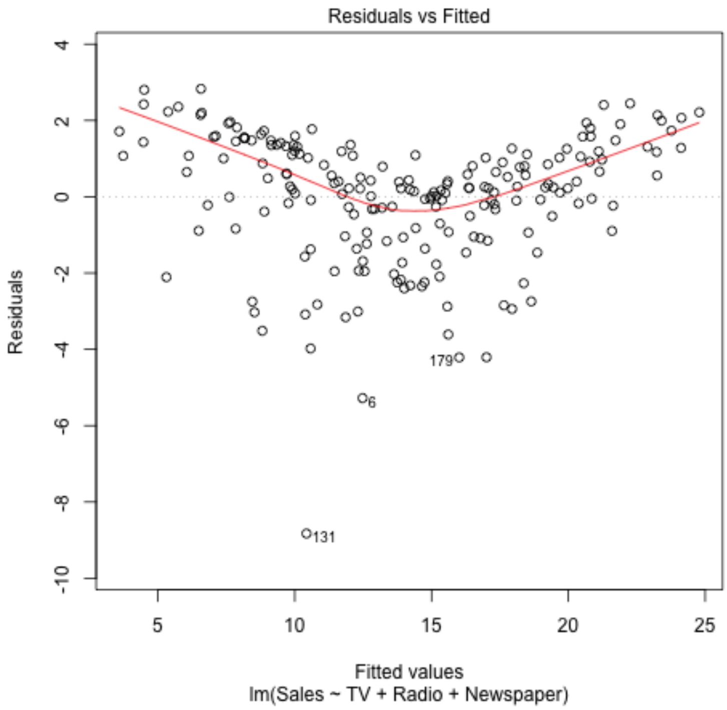
	TV	Radio	Newspaper	Sales
TV	1.00	0.05	0.06	0.78
Radio	0.05	1.00	0.35	0.58
Newspaper	0.06	0.35	1.00	0.23
Sales	0.78	0.58	0.23	1.00

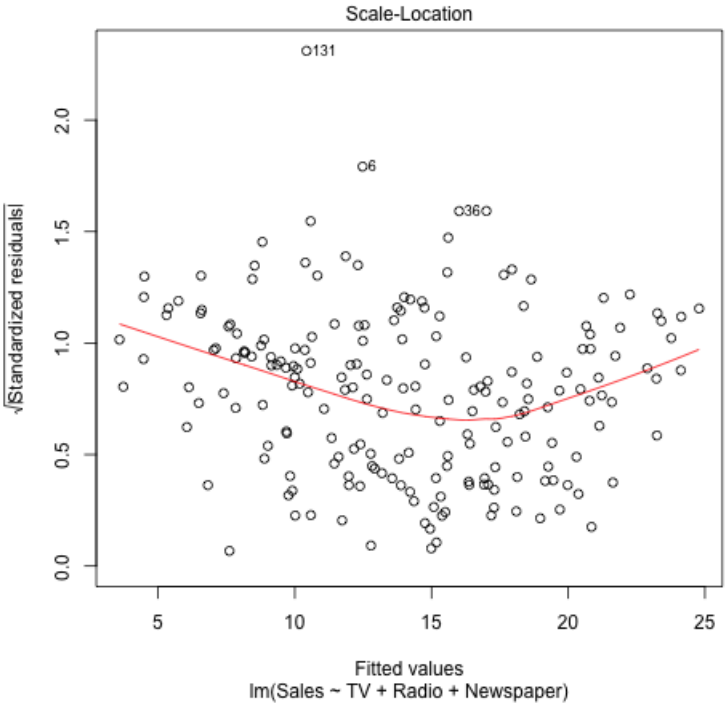
Table 5: Least Squares Coefficient Estimates from the Multiple Linear Regression of Sales on TV, Radio, and Newspaper

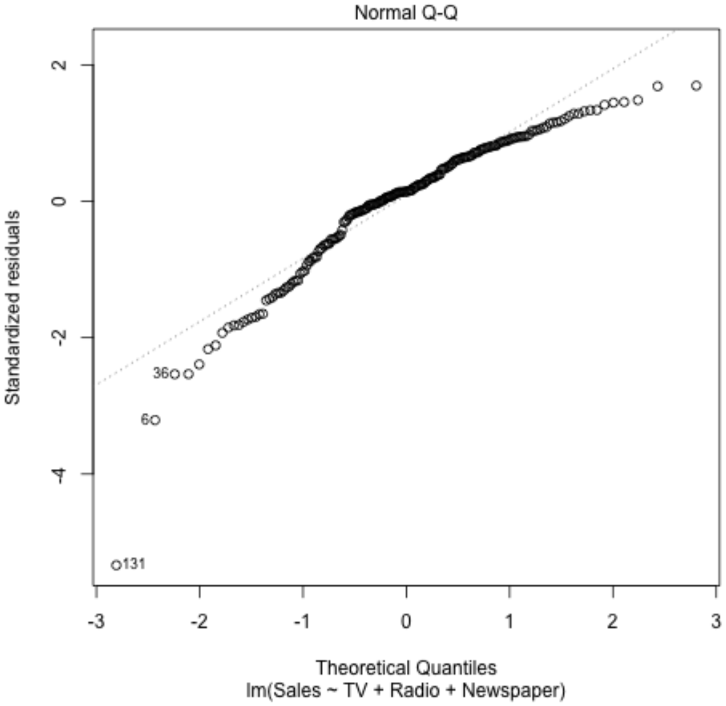
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.939	0.312	9.42	0.0000
TV	0.046	0.001	32.81	0.0000
Radio	0.189	0.009	21.89	0.0000
Newspaper	-0.001	0.006	-0.18	0.8599

Table 6: Quality Indices for the Multiple Linear Regression of Sales on TV, Radio, and Newspaper

	Quantity	Value
1	Residual standard error	1.69
2	R^2	0.90
3	F-statistic	570.00


Residuals vs Fitted
Residuals
4
2
0
-2
-4
-6
-8
-10
179
6
131
5
10
15
20
25
Fitted values
lm(Sales ~ TV + Radio + Newspaper)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
131
6
36
5
10
15
20
25
Fitted values
lm(Sales ~ TV + Radio + Newspaper)



From **Table 5**, we can see that the coefficient estimate for β_0 is 2.939, β_1 is 0.046, β_2 is 0.189, and β_3 is -0.001. The coefficient estimate for TV is about the same, and still remains statistically significant at the 1% and 5% level, since its p-value is NA and therefore less than .01 and .05. The coefficient estimates for Radio and Newspaper, however, have changed quite a bit from their estimated values in the simple linear regression. Radio remains statistically significant at the 1% and 5% level, since its p-value is NA and therefore less than .01 and .05. Newspaper, however, is no longer statstically significant because its its p-value is NA and therefore drastically higher than .01 and .05.

From **Table 6**, we can see that on average, the observed Sales data differs from its corresponding predicted value from the multiple linear regression model by 1.69, which translates into a difference of 1690 units of product sold. The R^2 value tells us that 89.7% of the variability in Sales is explained by our model–which provides evidence that the model fits the data well. The $F - statistic$ of 570 is not close to 1 at all, we can reject the null hypothesis that all there is no relationship between Sales and TV, Radio, and Newspaper.

Conclusions

Overall, we can see that there is indeed a difference in results when it comes to performing 3 simple linear regressions versus 1 multiple linear regression to capture the relationship between Sales and TV, Radio, and Newspaper. In fact, we were able to see that the coefficient estimate for Radio changed quite a bit through multiple linear regression and that the coefficient estimate for Newspaper was no longer statistically significant. While this casts some doubt on the usefulness of Radio, and Newspaper in predicting Sales (the response), the fact that the coefficient estimate for TV changed very little between models and remained statistically significant suggests that there is at least one useful predictor in the model. We also saw from the calculated R^2 value of 0.897 that the multiple linear regression model does, indeed, fit this dataset well. The fact that it is relatively close to 100% suggests that the prediction is reasonably accurate.