

# Hw2 - Simple Regression Analysis

Jacqueline Liu

October 31, 2016

## 1 Abstract

Linear regression is a simple, yet powerful method of modeling and supervised learning. Many real-world phenomena can be represented as a linear model, or with other models that employ linear regression. The well-known and well-referenced text **An Introduction to Statistical Learning** covers this technique in chapter 3, section 1 *Simple Linear Regression*. Here we will reproduce their results to highlight key steps and ideas.

## 2 Introduction

There is a wide variety of applications for linear regression; it can be used for prediction of response to a medical treatment, forecasting prices of fruits a month from now, or determining whether or not sleep and grades are correlated. Ultimately, the goal is to describe some trend or relationship in a dataset. Once the appropriate model is found, predictions can be made for certain variables if given the values of others and the strength of relationships can be assessed. Here, we'll be looking at a commercial setting and determining the relationship between advertising spending and sales of a given product. Specifically, we want to predict sales given the advertising budget for three different media outlets.

## 3 Data

The 'Advertising' dataset was 'collected' from 200 different markets. It records the advertising budgets (in thousands of dollars) for 'TV', 'Radio', and 'Newspaper' as well as the number of 'Sales' (in thousands of units) for each of the markets. Because 'Sales' is what we're ultimately trying to predict, it is our response variable while the other three columns will be our features.

## 4 Methodology

We're using linear models, which can be written as:

$$Y = \beta_0 + \beta_1$$

Given X and Y, the coefficients  $\beta_0$  and  $\beta_1$  are chosen to minimize the sum of squared errors:

$$SSE = \sum (Y - \hat{Y})^2$$

giving our regression line the name "least squares regression". There are other loss metrics we could use (weighted errors, least absolute deviation, L1-norm penalty) but least squares is one of the most widely used. We can regress 'Sales' on 'TV', then on 'Radio', and lastly 'Newspaper' to find three different linear models.

## 5 Results

The regression coefficients are:

```
> xtable(tv.sum$coefficients, caption="Sales onto TV")
```

	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: Sales onto TV

```
> xtable(radio.sum$coefficients, caption="Sales onto Radio")
```

	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: Sales onto Radio

```
> xtable(news.sum$coefficients, caption="Sales onto Newspaper")
```

	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: Sales onto Newspaper

The quality indices are:

```
> xtable(data.frame(Quantity=c('RSS', 'R2', 'F-stat'),
+                      Value=c(tv.sum$sigma, tv.sum$r.squared, tv.sum$fstatistic['value'])),
+       caption="TV Regression Quality Indices")
```

	Quantity	Value
1	RSS	3.26
2	R2	0.61
3	F-stat	312.14

Table 4: TV Regression Quality Indices

```
> xtable(data.frame(Quantity=c('RSS', 'R2', 'F-stat'),
+                      Value=c(radio.sum$sigma, radio.sum$r.squared, radio.sum$fstatistic['value'])),
+       caption="Radio Regression Quality Indices")
```

	Quantity	Value
1	RSS	4.27
2	R2	0.33
3	F-stat	98.42

Table 5: Radio Regression Quality Indices

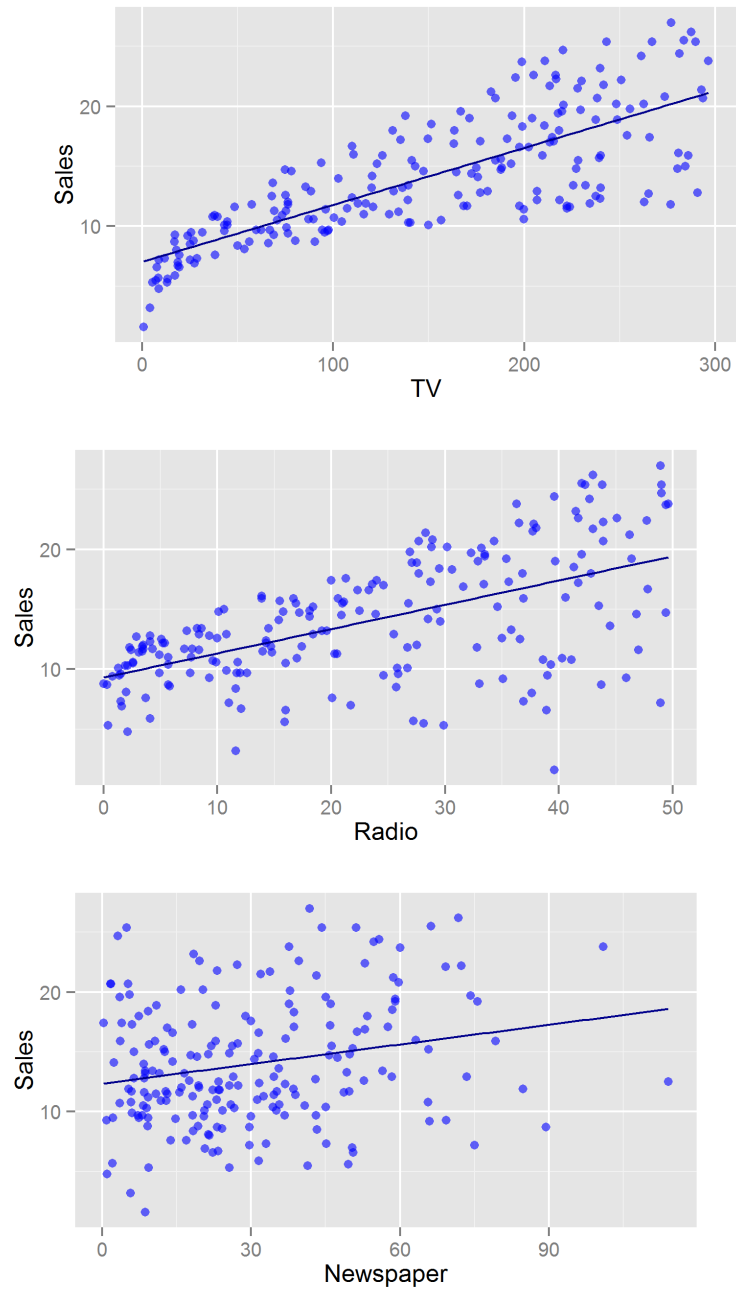
```
> xtable(data.frame(Quantity=c('RSS', 'R2', 'F-stat'),
+                      Value=c(news.sum$sigma, news.sum$r.squared, news.sum$fstatistic['value'])),
+       caption="Newspaper Regression Quality Indices")
```

	Quantity	Value
1	RSS	5.09
2	R2	0.05
3	F-stat	10.89

Table 6: Newspaper Regression Quality Indices

(NOTE: the p-values are not exactly 0, but so small that for all intents and purposes they basically are).

To visualize the data and the proposed model, here are scatterplots with the calculated regression lines:



## 6 Conclusion

Looking at these results, we can see though the intercepts and slopes are significantly different from 0 (meaning there is a relationship between the advertising budges and 0). However the slopes themselves are not all very large; only **Radio** has a slope in the tenths, vs hundreths, decimal place. However, this isn't to say that the company should spent all its advertising budget on radio advertisements. **TV** has the greatest  $R^2$ -value, meaning it is best at explaining the variation in **Sales**, which can be clearly seen in the scatterplots. While **Newspaper** has no obvious trend and all values of **Radio** are still correlated low **Sales** values, **TV** is the only to show increases in **Sales** consistently. Another factor to consider is how these features work together, a higher-dimensional relationship not captured in this analysis.