TV versus Sales Linear Regression Model

Lauren Hanlon

October 7, 2016

Abstract

For this project I will refer to the book *An Introduction to Statistical Learning* (by James et al).

My goal for this project was to conduct a simple analysis using the Advertising.csv dataset to look at sales across various products as a function of advertising budget, specifically for TV media spend.

In particular, we were answering the question: *Is there a relationship between TV advertising budget and sales?*.

Throughout this report I will conclude that there is a direct positive correlation between increased budget on TV media spend and sales.

Introduction

In order to answer the question if there is a relationship between the TV advertising budget and sales, we need to find concrete evidence of an association between TV media spend and sales. To find this association, I will use a *simple linear regression model*, which predicts a quantitiatve response Y on the basis of one predictor variable *X*. We can model this linear relationship as Y = β_0 + B1X. In particular, for this analysis looking at TV media spend versus sales, since we are trying to predict sales based off of TV media spend, our predictor variable, X, will be TV media spend, and our Y variable will be sales.

`sales` = β_0 + β_1 x `TV`

β_0 represents the *y-intercept*, which simply means that if X = 0 then what would our Y look like (i.e. if our TV media spend is \$0 what would our sales be). β_1 represents the *slope*, which is the amount by which Y increases/decreases as we change our X value. (i.e. if our TV media spend increases from \$100 to \$200, the change in sales would represent the slope)

How do we know what β_0 and β_1 are? The answer is: we don't, and this is the goal of a linear regression model. To find the coefficient estimtaes β_0 and β_1 such that the model fits the data. Think of a scatterplot with all the possible combinations of (TV spend, sales), then the linear model will find a *line of best fit* such that the line approximates the data as closely as it can by using a method of *minimizing the least squares.*

Data

The Advertising data set is 200 x 5 in dimensions. There are 200 rows, each row being a unique item and there are 5 columns:

- X: index
- TV: Advertising budget on TV (in thousands \$)
- Radio: Advertising budget on radio (in thousands \$)
- Newspaper: Advertising budget on newspaper (in thousands \$)
- Sales: Product sales (in thousands \$)

The table contains `sales` in thousands of units for a particular product as a function of advertising budgets (in thousands of dollars). In this specific analysis, we will use the `TV` and `sales` columns to look at the relationship between TV and sales.

Methodology

My methodology for this project was to first compute some simple statistics, to get an idea of how the data was formatted and spread out. Then I created a visual representation for the TV data as well as the Sales data, to see the distribution. I then created a linear model and plotted TV versus Sales to get an idea of their correlation.

Simple Summary

```
##
## Please cite as:
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

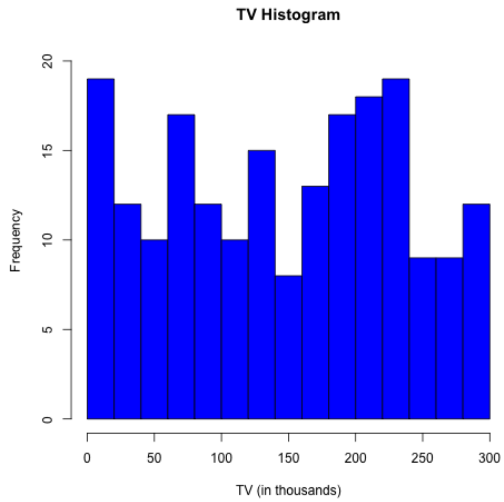
Table 1: Summary Statistics for TV spend and Sales (in thousands)

	V1
1	TV Sales
2	Min. : 0.70 Min. : 1.60
3	1st Qu.: 74.38 1st Qu.:10.38
4	Median :149.75 Median :12.90
5	Mean :147.04 Mean :14.02
6	3rd Qu.:218.82 3rd Qu.:17.40
7	Max. :296.40 Max. :27.00

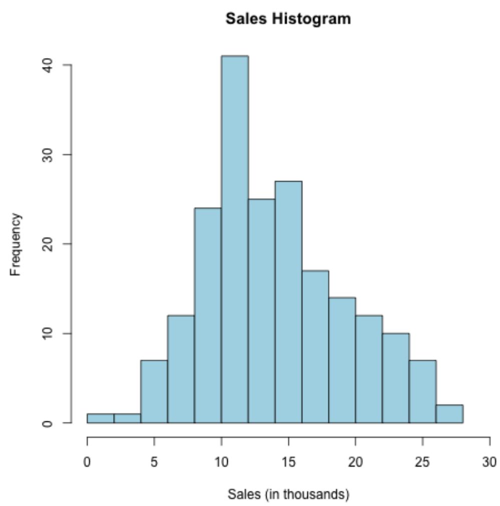
From this summary table, we can see that the spread of TV media spend went from 0.70 to 296.40 (in thousands), while sales spread from 1.60 to 27.00 (in thousands). This goes to show that there is greater standard deviation in TV media spend than in sales. The mean of TV fell at 147.04 with the median being 149.75, and the mean of Sales fell at 14.02, with the median at 12.90, which leads us to believe that they are similar in distribution. Let's look at this graphically to confirm this.

Histograms

I wanted to take a quick look at the data, to see how it is distributed. To do this I created histograms of the data to look at their frequency distribution.



From the histogram of TV we can see that it is indeed *not* normally distributed, and actually varies quite a lot

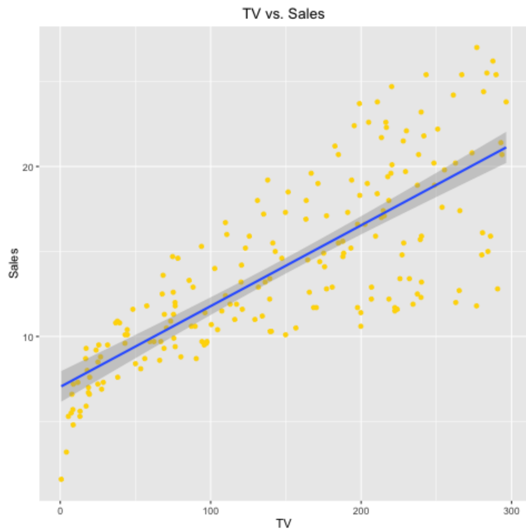


The histogram of sales actually shows that it has more of a normal distrubtion than the TV distribution, but tails more to the right whereas it drops quite steep on the left side, which a giant spike in the 10-13 range.

Scatterplot and linear model

Table 2: Linear Regression TV versus Sales

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.033	0.458	15.360	0
TV	0.048	0.003	17.668	0



This scatterplot represents the least squares fit for the regression of `sales` onto `TV`. Each blue point represents a pair (`TV`, `sales`), and the yellow line represents the linear model that fits the data. For each value of TV spend, we can now predict the sales if we were to spend that much on TV.

Diving deeper into the analysis, we can conclude that our y-intercept is 7.03259 and our slope is 0.04754.

Our formula for TV versus Sales is now:

`sales` = 7.03259 + 0.04754 x `TV`

More generally, this means that there is a +0.04754 correlation between an increase in TV spend and sales.

Results

The results show a direct positive correlation between TV media spend and sales. The result of this simple linear regression model would lead one to encourage the increase of TV media spend to improve sales. However, this comes with a few stipulations that you would want to check before jumping to increase your TV budget. For example, we would want to look at the correlations between radio and newspaper media spend as well to see if there might be a stronger correlation. If so, we would want to invest in those areas before investing more in TV. Another iteration on this linear model would be a more complex model. As you can see in the scatterplot, as TV spend increases, the distance between the points and the residual line increase, in both directions. This leads me to believe that there is some other factor involved as TV spend increases that either significantly increases, or decreases the impact of sales.

Conslusions

While the goal of this analysis was to find a relation between TV media spend and sales, the real purpose of this project was to determine the reproducibility of the analysis in James' *An Introduction to Statistical Learning*. Since our results are the same as his, we can conclude that his findings are indeed reproducible.