Report - Homework 3 Stat 159

Kartikeya Gupta

October 14, 2016

Abstract

This is Homework 3 for Stat 159 - RCDSC. For this homework, we are reproducing the main results displayed in Section 3.1, Chapter 3 of the book An Introduction to Statistical Learning.

Introduction

Through this report, we try to understand the relationship between Sales, TV, News and Radio advertisements, and come up with a simple model that can be used to understand such a relationship. This would help us determine whether the above mentioned media ads are actually successful at increasing sales or not and if yes, then what is their impact.

Data

The data set we are working with is called the "Advertising"" data set. It consists of the Sales of a particular product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, Radio, and Newspaper.

Sales are in thousands of units.
Advertising budgets for TV, Radio and Newspaper are measured in thousands of dollars.
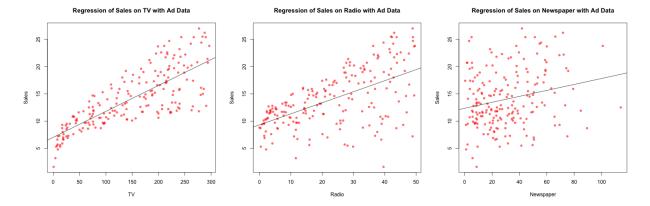
Methodology

In this paper, we will be using a Multiple Linear Regression model which is used to find a quantitative relationship between the independent variables TV, Newspaper, Radio and the dependent variable Sales. Here the is the equation:

$$Sales = \beta_0 + \beta1 TV + \beta2 Radio + \beta3 Newspaper + \epsilon$$

Here the β_0 is the constant term and the β_1 is the coefficient of TV Ads. Note we fit the regression by using the Ordinary Least Squares Estimator method.

Results

To be able to visualize the data we create the scatterplot given below. through this scatterplot, we can see that there are various values of of Sale-vs-Tv and an upward sloping line. This line indicates a positive linear relationship between Tv ads and Sales. Further, we notice that the graph is heteroscadastic which means that the variance is changing given the slope. Initial sales have a small variance as compared to the large variance towards the right end of the graph where TV ad budget is far greater. This could signify a rather loose casuality when the budget increases alot. Note that the line going through the plot is essentailly the fitted line of the regression.

Going deeper into this analysis, we focus our direction to the model and its quantitative properties.

First we have the coefficients from this fitted regression given as below along with their std. errors and t-statistics.

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

Table 1: Regression Output

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

Table 2: Regression Output

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

Table 3: Regression Output

From the intercept coefficient, we can see that if there was a $0 ad budget then we would have 7.0325*1000 ~ 7032 units of sale. This represents the constant coefficient.

Moving on to the TV coefficient, we can understand that for every unit increase in TV budget the Sales would be impacted by +0.0475 units. Which means that if we increase the add budget by $1000 then we would ideally see an increase in the sales by 47.5 units. This outlines the models ability to find a relationship between TV ads and actual Sales. Also, notice that the t-stat on the TV coefficient is 17.67 which indicates that this result is highly statistically significant.

From this, we know that allocation to the TV budget is actually efficient at increasing sales of the product. The model also helps in the prediction of the sales given a certain TV budget. This could help make future decisions about how much to allocate for marketing campaigns focused on TV.

Finally, to test whether this model is actually good we measure a few more variable. Please look at the table below.

The first term "Residual Standard Error" measure the lack of fit of the model. This means that on average how much would the variance be from the given estimates. Here the value of 3.26 indicates that given the tv advertising the sales could be off by 3,2600 units on average. This error would be lower if the model had less variance which could make our estimate prediction better.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9389	0.3119	9.42	0.0000
TV	0.0458	0.0014	32.81	0.0000
Radio	0.1885	0.0086	21.89	0.0000
Newspaper	-0.0010	0.0059	-0.18	0.8599

Table 4: Regression Output

	Quantity	Value
1	Residual Standard Error	1.69
2	R2	0.90
3	F-Statistic	564.45

Table 5: Measure of Fit Statistics

The second term "R^2" is another measure of fit statistic. It describes the model in the form of proportions of variance in data explained by the model. Here 0.612 means that 61.2% of the variance can be explained by this model.

Finally, the F-statistic is a measure of the statistical significance of the model. A value of 312.1 is high enough that we can say that the results are highly statistically significant.

Conclusion

Through this paper, we learn that TV ads play a significant role in increasing poduct sales. We also learn about the approximate quantitative difference tv ads made on sales for this given data. The results are statistically significant and 61.2% of the variation can be explained by our linear model. This analysis can be helpful for anyone who is trying to understand the impact of tv ads dollars and how to allocate budgets accordingly. Finally, this is precisely why we see ads on tv. If it never increased sales then there would be no point of showing the advertisements.

	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 6: Correlation Table