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 and TV, News, 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

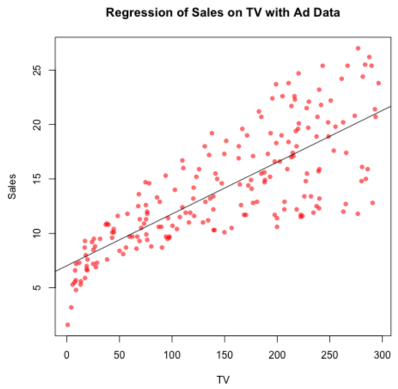
In this paper, we will be using a Multiple Linear Regression model which is used to find a quantitative relationship between the independent variables TV, Newspaper, Radio and the dependent variable Sales. Here is the equation:

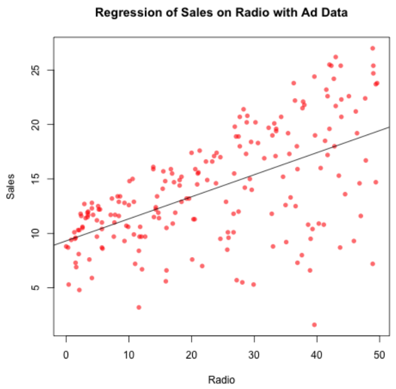
$$Sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper + \epsilon$$

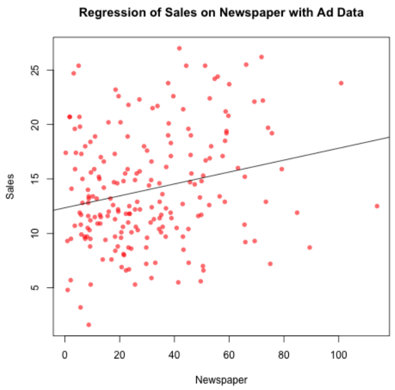
Here the β_0 is the constant term. β_1 is the coefficient of TV Ads. β_2 is the coefficient of Radio Ads. And β_3 is the coefficient of Newspaper Ads. ϵ is the error term. Note we fit the regression by using the Ordinary Least Squares Estimator method.

Results

To be able to visualize the data we create the scatterplots given below. Through these scatterplots, we can see that there are various values of Sale-vs-Tv and an upward sloping line. We can observe a similar relations betwen Sales-vs-Radio/Newspaper though the dots are more scattered and the correlation is positive but lower. The lines on the plots indicate a positive linear relationship between Tv, Newspaper, Radio ads and Sales. Further, we notice that the graphs are 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 and Radio ad budgets are far greater. However, for newspaper there isnt alot of data for high budgets maybe because its cheaper. However, the effectiveness for the same budget can vary alot - as seen on the left half of the plot. In general, for Tv and Radio, this could signify a rather loose casuality when the budget increases alot. But for newspaper, we see little correlation between budget and sales. Note that the lines going through the plots are essentailly the fitted lines of the regression.







Going deeper into this analysis, we focus our direction to the single linear regression 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 of Sales on TV

	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 of Sales on Radio

	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 of Sales on Newspaper

From the above:

- TV - We can see that if there was a $0 ad budget then we would have 7032.5935491 units of sale. This represents the constant coefficient. For every $1000 increase in TV budget, there would be an additional 47.5366404 units of sale. Also, notice that the t-stat on the TV coefficient is 17.6676256 which indicates that this result is highly statistically significant.
- Radio - We can see that if there was a $0 ad budget then we would have 9311.6380952 units of sale. This represents the constant coefficient. For every $1000 increase in Radio budget, there would be an additional 202.4957834 units of sale. Also, notice that the t-stat on the Radio coefficient is 9.9207655 which indicates that this result is highly statistically significant.
- Newspaper - We can see that if there was a $0 ad budget then we would have 12.3514071 units of sale. This represents the constant coefficient. For every $1000 increase in Newspaper budget, there would be an additional 54.6930985 units of sale. Also, notice that the t-stat on the Newspaper coefficient is 3.2995907 which indicates that this result is not statistically significant thus we cannot conclude that there is a strong relationship.

	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

From this, we know that allocation to the RAdio budget is actually most efficient at increasing sales of the product, while TV budget is still quite good. There is a rather insignificant difference made with newspaper ads and they should probably not be continued. This model also helps in the prediction of the sales given a certain TV+Radio+News budget. This could help make future decisions about how much to allocate for marketing campaigns.

From this table of Correlations,

	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: Correlation Table

we can see that TV-Sales and Sales-Radio have high positive correlations while the Sales-Newspaper term has a low but positive relationship.

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

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

Table 6: Measure of Fit Statistics

The first term "Residual Standard Error" measures 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 1.6941763 indicates that given the tv advertising the sales could be off by 1694.1762936 units on average. This error would be lower if the model had less variance which could make our estimate prediction better.

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 89.7210638 means that 89.7210638% 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 564.4516148 is high enough that we can say that the results are highly statistically significant.

Conclusion

To directly answers the questions on Page 75 on ISLR:

1. Is at least one of the predictors useful in predicting the response?
   Yes, TV and Radio are quite helpful predicting sales.

2. Do all predictors help to explain the response, or is only a subset of the predictors useful?
   No, Newspaper is not significant enough to be conclusive.

3. How well does the model fit the data?
   The model captures 89.7210638% of the variance of the data, which is actually quite good.

4. How accurate is the prediction? The accuracy of the prediction is a function of the error term, the bias in the model and the fact that this is a prediction and not an actual value. Due to this, we look at the confidence intervals of the perdictors and the confidence intervals of the sales. This would give us a better idea about how accurate we can be on average.

In conclusion, through this paper, we learn that TV and Radio ads play a significant role in increasing product sales. We also learn about the approximate quantitative difference tv and radio ads made on sales for this given data. We also find that newspaper ads are actually not useful for this prediction model. Their variation for any given point is too much to be able to find a concrete answer about their effectiveness. We notice that the results are statistically significant and majority 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 and radio ads dollars and how to allocate budgets accordingly. Finally, this is precisely why we see ads on tv and hear them on radio. If it never increased sales then there would be no point of paying for those advertisements.