HW 2: Simple Regression Analysis

Todd Vogel

10/6/2016

Abstract

The following reports looks into data on advertising costs across multiple media including television, radio, and newspaper as well as sales. By utilizing linear modeling we effectively establish the relationship between advertising and sales numbers, specifically focusing on television. The results are a reproduction of the work in Chapter 3.1 in the book *An Introduction to Statistical Learning*.

1 Introduction

This projects looks to isolate the effects of television advertising on a companies sales. By focusing on these two attributes and applying a linear model, we can begin to understand how the two variables are related. If the association is apparent, the linear model created will allow us to effectively predict increases in sales from changes in television advertising in the future. Additionally, this process can be applied to newspapers and radio to determine their impact as well.

2 Data

The data utilized in the project is exclusively an Advertising data set with 4 variables: TV, $Radio$, $Newspaper$, and $Sales$. $Sales$, our dependent variable is measured in thousands of units and TV, $Radio$, and $Newspaper$ are our independent variables measured in thousands of dollars. Each data point, or row, represents an individual product (of which there are 200, in as many different markets).

3 Methodology

The completion of this project required many steps and methodologies. However, the most important is regression. To determine the relationship between television and sales we applied the linear model:

$$Sales = \beta_0 + \beta_1 * TV \tag{1}$$

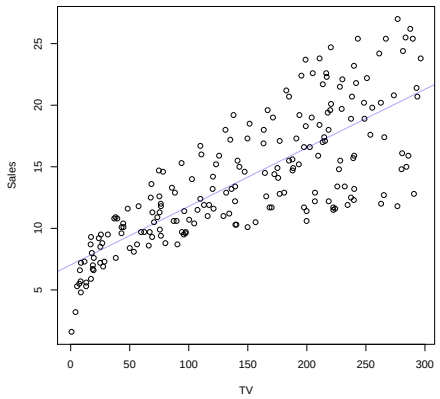

Figure 1: Sales v. TV Scatterplot and Regression Line

Then, using the lm function in R, we calculate the values for the coefficients β_0 and β_1 (through the Ordinary Least Squares calculation). The remainder of the project was constructed using command line and makefile in order to ensure reproducibility. The file and directory structure was created in the terminal and r scripts were intertwined to create a reproducible output with Makefile and the *make* function.

4 Results

Table 1: Regression Coefficients and Quality Indices Not able to load R table into LateX document (Sindhuja said it was fine to leave out)

First, this data presents a statistical breakdown of the residuals. In this case a residual is the error of the resulting prediction from each data point (difference between actual and predicted value). This data then provides the values of the coefficients β_0 and β_1 as 7.03 and 0.05 respectively. Additionally the RSE, R^2, and $F - statistic$ are 3.26, 0.61, and 312.1 respectively. Because the RSE is close to 0 and relatively small compared to sales numbers our model is a good fit.

This scatterplot looks at each of the 200 data points and graphs the sales value on the y axis and the television advertising value on the x axis. Looking at the graph, there is clearly a strong direct relationship between increasing TV ads and increasing sales. However, it is important to not that the distribution of the data is not homoscedastic as the residuals about the regression line increase with more advertising.

5 Conclusions

In the end, it is plausible to conclude that the relationship between television advertisement and sales is directly related. Further, RSS and other indicators are low meaning that the linear model calculated is an effective predictor of sales using data from television advertisement. This realization sets the precedent for greater modeling with the advertising data. It would now be useful to determine the relationship between newspapers and sales and radio and sales.