

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 modelling 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**.

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.

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).

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:

$$Sales = \beta_0 + \beta_1 TV$$

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.

Results

Table 1: Regression Coefficients and Quality Indices

```
##
## Call:
## lm(formula = Sales ~ TV, data = advertising.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.032594   0.457843   15.36  <2e-16 ***
## TV           0.047537   0.002691   17.67  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Table 2: TV and Sales Statistics

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
tv.statistics	0.7	74.38	149.8	147.00	218.8	296.4
sales.statistics	1.6	10.38	12.9	14.02	17.4	27.0