Multiple Regression Analysis

Bret Hart

October 14, 2016

Abstract

The aim of this report is to reproduce some of the graphical, statistical, and tabular results displayed in chapter 3.2, *Multiple Linear Regression*, of **An Introduction to Statistical Learning**. Referred to as **ISLR**, the textbook is a manifesto of Machine Learning and Linear Models, teaching the material in an approachable yet sophisticated way. In addition, the data used to generate all of the graphs, plots, etc. in the text are freely available - advancing and standing for the tenants of reproducible research, even in a textbook. We seek to create an automated repository which can recreate the findings that they display, using the same data set.

Introduction

The data set which we are studying is an Advertising data set - it is a collection of money spent in 200 different markets on Advertising, with each market's corresponding Sales figures. The data includes information on Television, Radio, and Newspaper advertisement, and for this multiple linear regression project, we will be considering how all three play a role in determining and predicting their corresponding Sales figures. We would like to determine whether there is a meaningful, significant relationship between the three advertisement mediums and Sales. Using these results, we would like to be able to predict future Sales figures based on potential amounts of Advertisement expenditure. Ultimately, we would like to make sophisticated, informed decisions on how to form an Advertising plan in the future, with all of the three possibly confounding variables considered simultaneously and in relation to one another. We want to model this relationship effectively and correctly, and use the model to predict future sales and create a profitable Sales plan.

Additionally, we seek to reproduce certain pertinent figures and tables shown in chapter 3.2 of **ISLR**, intermediary statistical explanations which help us to better understand all the ways in which the predictors and dependent variable are related. The tables, thus, are valuable both because making them proves the reproducibility of the **ISLR** text, but also because the tables do truly help us to understand the Advertising data set better.

Data

More specifically, the Advertising data sets contains **Sales** (in thousands of units) of a particular product in 200 different markets, supplemented by advertising budgets (in thousands of dollars) for the products in three different forms of media: **TV**, **Radio**, and **Newspaper**. For this project we are going to focus on the relationship between the three collected predictors, **TV**, **Radio**, **Newspaper**, and their response, **Sales** - for the purposes of specifically reproducing the figures and findings in **ISLR** and to better understand multiple linear regression.

Methodology

As stated previously, we are focusing on the three advertising mediums of **TV**, **Radio**, and **Newspaper** and their relationship with **Sales**. We will consider both their individual, 1-1 relationship with **Sales**, as well as their combined predictive effect. To do this, we will assume and use the multiple linear model:

$$\textbf{Sales} = \beta_0 + \beta_1\textbf{TV} + \beta_2\textbf{Radio} + \beta_3\textbf{Newspaper}$$

To estimate the coefficients β_0, β_1, β_2, and β_3, we fit a regression model via the multiple least squares criterion.

Results

First, we estimate the correlation coefficients and linear relationship between each of the three individual predictors and **Sales**, carrying out a simple linear regression for **Sales** on each of the three predictors.

Table 1: Simple Regression of Sales on TV

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.033	0.458	15.360	0.000
TV	0.048	0.003	17.668	0.000

Table 2: Simple Regression of Sales on Radio

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.312	0.563	16.542	0.000
Radio	0.202	0.020	9.921	0.000

Table 3: Simple Regression of Sales on Newspaper

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.351	0.621	19.876	0.000
Newspaper	0.055	0.017	3.300	0.001

However, these simple regression coefficients and their resultant p-values can be deceiving - they do not necessarily imply that the coefficients will be the same in the multiple linear model, or even that the p-values for each predictor are truly significant when considered with the other predictors in mind. Thus, we then create a table of the coefficients and their significance in the collective multiple linear regression model to gain a better sense of the role each predictor plays in determining **Sales**.

Table 4: Multiple Linear Regression of Sales on TV, Radio, and Newspaper

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9389	0.3119	9.4223	0.0000
TV	0.0458	0.0014	32.8086	0.0000
Radio	0.1885	0.0086	21.8935	0.0000
Newspaper	-0.0010	0.0059	-0.1767	0.8599

To gain further understanding into the relationships between the four variables in the Advertising data set, we can examine all of their respective correlations to one another. This can perhaps shed insight into how each predictor interacts and relates to the others, possibly in a confounding manner. Of course, we can also see each predictor's respective correlation with **Sales**, which is beneficial in its own right.

Table 5: Correlation matrix for the four Advertising variables

	TV	Radio	Newspaper	Sales
TV	1.0000	0.0548	0.0566	0.7822
Radio	0.0548	1.0000	0.3541	0.5762
Newspaper	0.0566	0.3541	1.0000	0.2283
Sales	0.7822	0.5762	0.2283	1.0000

Now that we have considered each of the predictors individually, and their relation to one another, we can start to consider the multiple regression model, of form listed above. This model is near impossible to graph in

a human-readable format without some other outside tools, so we cannot simply examine the regression model and its graph to gain insight into the model. Although the model may not have a simply graphable form, that importantly plays little into its ability to actually test different values for each of the three advertising amounts - the model will still spit out numeric results, regardless of feasibility of visualization!

Let's start with a table of some basic statistics: the multiple Residual Standard error (RSE), the multiple R^2, and the F-Statistic.