Multiple Regression Analysis

Yuyu Zhang

Abstract

In this report we extend the *simple regression analysis* from last time and try to reproduce the main results displayed in section 3.2 *Multiple Linear Regression* of the book **An introduction to statistical Learning**

Introduction

The goal is to extend the analysis of advertising data to see, besides TV, if other two medias newspaper and radio have an effect on the sales of a particular product. And if so, develop an accurate model that can be used to predict sales on the basis of three media budgets.

Data

The Advertising data set consists of the **Sales** (in thousands of units) of a particular product in 200 different markets, along with advertising budgets (in thousands of dollars) for the product in each of those markets for three different media: **TV, Radio**, and **Newspaper**.

Methodology

The multiple linear regression model gives each predictor a separate slope coefficient in a single model:

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

To estimate each coefficient we fit a regression model via the least square criterion. For RSE, R squared and F-statistics we used the formulas in the book to carry computation.

Results

The simple linear regression model for each media is shown in the tables below:

Table 1: Simple Regression of Sales on TV

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.03	0.46	15.36	0.00
TV	0.05	0.00	17.67	0.00

Table 2: Simple Regression of Sales on Radio

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.31	0.56	16.54	0.00
Radio	0.20	0.02	9.92	0.00

Table 3: Simple Regression of Sales on Newspaper

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.35	0.62	19.88	0.00
Newspaper	0.05	0.02	3.30	0.00

Coefficient estimates of the least squares model is shown below:

Table 3:Coefficient estimates of the least squares model

	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

Correlation matrix of three medias and sales is summarized in the table below:

Table 4: Correlation matrix of TV, radio, newspaper and sales

	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

Lastly, RSE, R square and F-statistic of the least square model is summarized in the table below:

Table 5:Other Relationship Statistics

	Quantity	Value
1	Residual standard error	1.6855
2	R squared	0.8972
3	F-statistic	570.2707

Answers to additional questions:

- When there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1. Since the F-statistics from multiple regression model is 570>1, it provides compelling evidence against the null hypothesis.In other words, the large F-statistic suggests that at least one of the advertising media must be related to sales.
- We use backward selection. First include all the variables and then inspect the p-value. We set the criteria for variable being relevant is when p-value is less than 0.01. From Table 3, it's obvious that newspaper exceeds the p-value threshold and is the least statistically significant. Therefore, we say that only TV and Radio have observable effects on sales.
- Two of the most common numerical measures of model fit are the RSE and R^2. The model that uses all three advertising media to predict sales has an R2 of 0.8972. On the other hand, the model that uses only TV and radio to predict sales has an R2 value of 0.89719. In other words, there is a small increase in R2 if we include newspaper advertising in the model that already contains TV and radio advertising.. The fact that adding newspaper advertising to the model containing only TV and radio advertising leads to just a tiny increase in R2 provides additional evidence that newspaper can be dropped from the model.The model that contains only TV and radio as predictors has an RSE of 1.681, and the model that also contains newspaper as a predictor has an RSE of 1.686.In contrast, the model that contains only TV has an RSE of 3.26.This corroborates our previous conclusion that a model that uses TV and radio expenditures to predict sales is much more accurate (on the training data) than one that only uses TV spending.
- We would use both 95% confidence interval and prediction interval to quantify the uncertainty. For example, given that $100,000 is spent on TV advertising and $20,000 is spent on radio advertising in each city, the 95 % confidence interval is [10,985, 11,528]. On the other hand, given that $100,000 is

interval spent on TV advertising and $20,000 is spent on radio advertising in that city the 95 % prediction interval is [7,930, 14,580].Both intervals are centered at 11,256, but that the prediction interval is substantially wider than the confidence interval, reflecting the increased uncertainty about sales for a given city in comparison to the average sales over many locations

Conclusions

In conclusion, we used multiple regression models to visualize the relationship between three advertising medias and their effects on product sales. The analysis suggests that a model with TV and radio as predictors are the more suitable one while newspaper has negligible effects on the product sales