Multiple Regression Analysis

Junyu Wang

Oct 14th, 2016

Abstract

In this report, we reproduce the main results displayed in section 3.2 *Multiple Linear Regression* (chapter 3) of the book *An Introduction to Statistical Learning*.

Introduction

The overall goal is to provide insights about whether advertising through different tunnels improves sales. In this report, we look at how TV, Radio and Newspaper affect sales number. We apply multiple linear regression models between different advertising tunnels and sales, and therefore can use those models to maximize sales number with certain amount of advertising budget.

Data

The advertising dataset consists of *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*, *Newspaper* and *Radio*).

Methodology

We consider *Sales* and *TV*, *Radio*, *Newspaper* in our dataset and try to fit them in a multiple regression model:

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

In this equation, the coefficients β_0, β_1, β_2, β_3 represents the association between Sales and the three advertising tunnels repsectively, and e is the error term. In order to find the best regression model, we need to find the best coefficients that minimizes the sum of squared residual.

Results

We get the 3 simple linear regression models' coefficients in tables below:

	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: Linear 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: Linear 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: Linear Regression of Sales on Newspaper

From the table above, we can see that the three slops are 0.04753664 for TV, 0.2024958 for Radio, and 0.0546931 for Newspaper. And because Radio is has the steepest slope, we can see that if we were only allowed to invest in oen of the three media, Radio will benefit us the most.

However, most of the times the three medias are not completely separated, and they might correlate with each other to affect sales number. Therefore in order to gain insights into their correlation, we build a multiple linear regression model as described below.

	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: Multiple Linear Regression of Sales on TV, Radio, Newspaper

we can see that the coefficients between sales and TV, radio are statistically significant, and because the coefficients between Newspaper and TV has a large p-value, we can treat it as a sign of weak relation. And this can be further demonstrated with the correlation matrix below.

	X	TV	Radio	Newspaper	Sales
X	1.00000	0.01771	-0.11068	-0.15494	-0.05162
TV	0.01771	1.00000	0.05481	0.05665	0.78222
Radio	-0.11068	0.05481	1.00000	0.35410	0.57622
Newspaper	-0.15494	0.05665	0.35410	1.00000	0.22830
Sales	-0.05162	0.78222	0.57622	0.22830	1.00000

Table 5: Correlation Matrix

From the matrix we can see that the correlation bewteen Radio and Newspaper is 0.05480866, this shows that radio and newspaper are usually used together. Therefore when sales number increases as Newspaper advertising budget increases, this doesn't necessarily mean it's newspaper that boosts the sales number, instead it's because in the meantime Radio advertising budget also increases.

So the final question is "Is our multiple linear regression model reliable?" We can show this from the table below

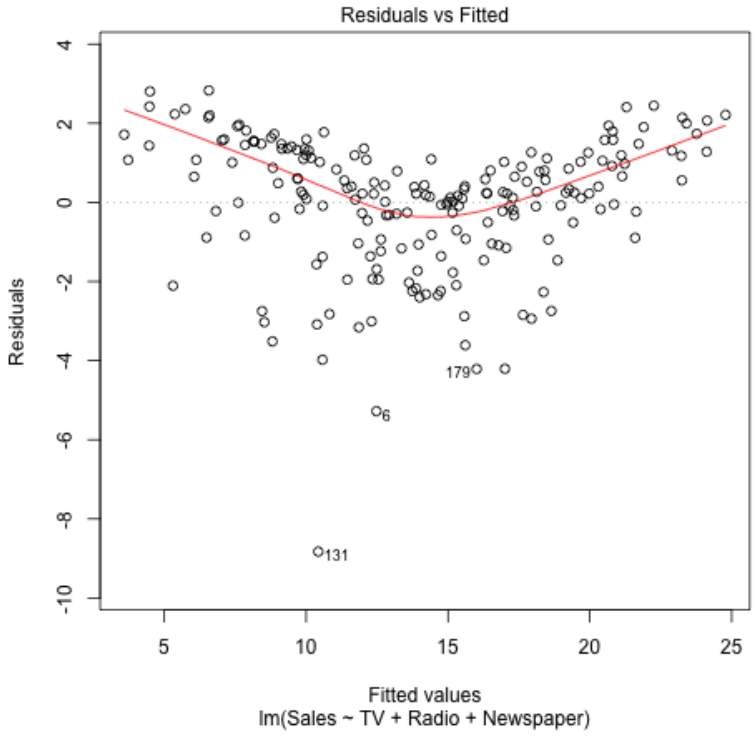
	col_names	col_data
1	Residual Standard Error	1.69
2	R-square	0.90
3	F-Statistics	570.27

Table 6: RSE, R-Squared, F-Statistics of Multiple Linear Regression Model

According to statistics shown above. RSE tells us usually the prediction error is 1.6855104 units; R-squared tells us 90% of changes in sales can be related to changing in advertising budgest, and this number is a really good sign of model reliability. If sales has nothing to do advertising, F-statistics should be close to 1, but in our statistics, F-statistics is 570.2707037 which is much greater than 1, this also shows strong correlation

between sales and advertising. Therefore, we can be sure that the relationship between sales and advertising exists.

This is scatterplot with Residual vs Fitted Value



Conclusions

After training both simple linear regression models and multiple linear regression models, we find out that multiple advertising tunnels are usually correlated and conclusion drawn simply from simple linear regression model is not inclusive. And from the RSE, R-Squared, F-Statistics of our multiple linear regression model we conclude that this model is reliable.