Multiple Regression Analysis

Aoyi Shan

Oct 12, 2016

Abstract

In this report, we summarize the steps we took toward replicating the results in Chapter 3, Linear Regression, from the book "An Introduction to Statistical Learning" by Gareth James, Deniela Witten, Trevor Hastie and Robert Tibshirani. In this project, we apply computational toolkits such as lm and summary function, graphic devices such as scatterplot matrix and residual plots and essential elements that enable a reproducible workflow to reproduce this multiple regression analysis.

Introduction

The main purpose of this project is to predict sales with 3 predictors, advertising budget in TV, newspaper and radio. Running a multiple regression model will provide us with great insights about the relationship between those variables and the regression coefficients represent the relative pairwise strength. After we run the multiple regression model, we can interpret key statistics such as slope, intercept, t-statistics and R^2 to determine the quality of the regression and offer advice on how to improve sales by effectively managing its advertising budget.

Data

The data set we used in the analysis can be downloaded from http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv. It contains four columns, sales and the advertising budgets for three different types of media, TV, newspaper and radio. TV represents the TV Advertising budget in thousands of dollars, and similar units are apply to newspaper and radio columns and Sales is the corresponding product sales in thousands of units.

Methodology

We start by setting up the multiple linear regression function with TV, newspaper and radio as independent variables and Sales as the dependent variable:

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

In this equatioin, β_0 is the intercept, each β_j for j > 0 quantifies the association between the regressor X_j and the response Y and e is the error term. In order to fit in a plane that is as close as possible to all the data points we have and minimize the residuals, the estimator for the slope and intercept should be obtained by running a regression model with the criterion of minimizing the sum of squared vertical distances between each observation and the plane.

Results

	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: Simple Regression of Sales on TV

	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 2: Simple Regression of Sales on Newspaper

	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 3: Simple Regression of Sales on Radio

Those 3 tables are the resulting coefficients when we fit a simple linear regression for each predictor. For every addition 1000 dollars we spend on TV advertising, on average, we expect to see 47.5 units increase in sales, while if we invest the same amount in newspaper or Radio, we expect to see an increase in sales by 54.7 or 202 units respectively. Therefore, we can conclude that radio advertising is the most efficient if we run three separate simple linear regressions.

However, this method ignores the effect of the other two media in predicting and it is highly possible that those three predictors are correlated. To obtain a better estimate, we extend the model by including 3 predictors and running a multiple regression to determine the coefficients.

	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 Regression of Sales on TV, Radio and Newspaper

Based on this regression output, for a given amount of radio and newspaper advertising, on average, investing an additional 1000 dollars in TV is associated with an increase in sales by 45.8 units. And in a similar manner, if we fix the budget for TV and newspaper, we predict that an additional 1000 dollars increase in Radio budget will lead to a 189 units increase in sales.

Besides, we can also see that the coefficient between sales and TV and radio are statistically significant, while it is a weak relationship between newspaper and sales since the coefficent comes with a large p-value. The correlation matrix provides us with an explanation.

	TV	Radio	Newspaper	Sales
TV	1.000000	0.054809	0.056648	0.782224
Radio	0.054809	1.000000	0.354104	0.576223
Newspaper	0.056648	0.354104	1.000000	0.228299
Sales	0.782224	0.576223	0.228299	1.000000

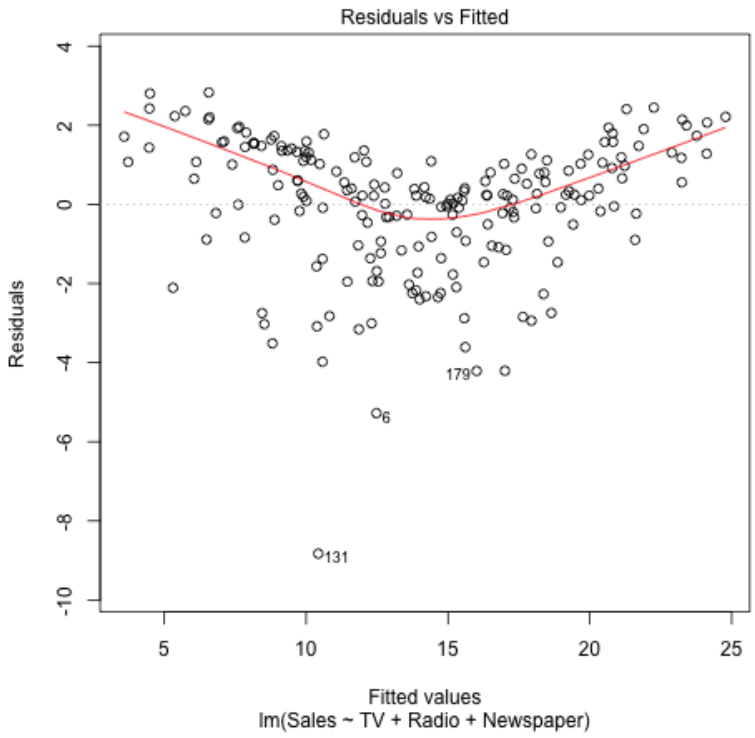
Table 5: Correlation Matrix

The correlation matrix indicates that the correlation between advertising budget in radio and newspaper is 0.354. This implies that those two medias are often used together. When the newspaper advertising budget increases, it singals that we also have a higher budget in radio advertising, and that is the factor leading to an increase in sales, not newspaper.

	Statistics	Value
1	Residual Standard Error	1.686
2	R-square	0.897
3	F-Statistics	570.3

Table 6: Regression Quality Statistics

We can see that Residual Standard Error is 1.686, which means the typical error the model is making in predicting the sales is 1.686 units. Then, $R^2 = 0.897$, which means that 89.7% of the variation in Sales can be explained by the change in advertising budget, which implies that our multiple regression model fits the data very well. F-statistics is used in the hypothesis testing process to determine whether all the coefficients are 0. If there is no relationship between the response and predictors, F-Statistics should be closer to 1. Here, since we have a large value for F-Statistic, we can conclude that there exists a relationship between sales and advertising budgets.



From the graph we can see that there is an obvious relationship between X and Y and the fitted multiple regression line seems to represent the relationship pretty well.

Conclusions

By running multiple regression with the data set the book provides, we are able to replicate the result and arrive at several conclusions:

1. Based on the multiple regression coefficient, only advertising budget in TV and radio contribute to predicting sales. As we explained with the correlation matrix, advertising in newspaper is correlated with the budget of radio, and doesn't help with promoting sales.
2. A large R^2 value indicates that the model fit the data very well since most of the variation in the response can be explained by the change in advertising budgets.
3. RSE represents the typical error the model is making in the prediction and a relatively small RES indicates that the model we come up with can achieve reasonable accuracy.