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 in 200 different markets.

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 and each β_j for j > 0 quantifies the association between the regressor X_j and the response Y. In order to fit in a plane that is as close as possible to the 200 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

From the table, we can see that the estimated slope is 0.05, which means that for every addition 1000 dollars we spend on TV advertising, on average, we expect to see 47.5 units increase in sales. The intercept can be interpreted as when we don't spend any money on TV advertising, we expect the sale to be 7032 units. We have large t-values and small p-values for both the slope and the intercept which imply statistical significance.

	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

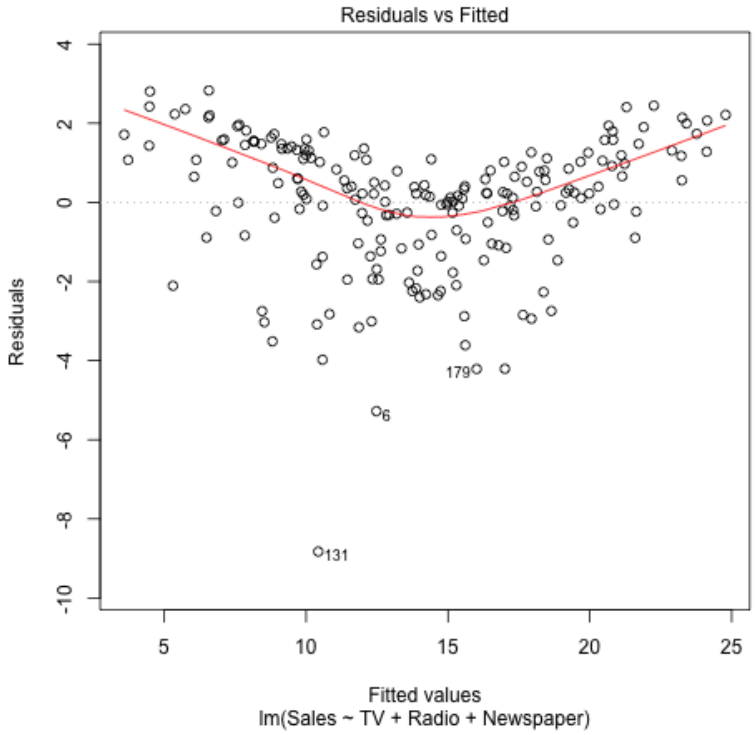
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.3514	0.6214	19.88	0.0000
Newspaper	0.0547	0.0166	3.30	0.0011

Table 2: Simple Regression of Sales on Newspaper

So based on the regression output, we can conclude that there exists a relationship between the advertising budget and sales.

% latex table generated in R 3.2.4 by xtable 1.8-2 package % Sun Oct 9 02:34:13 2016

We can see that $R^2 = 0.6119$, which means that 61.19% of the variation in Sales can be explained by the change in TV advertising. In addition, correlation coefficient equals the square root of R^2 which is 0.78. This implies a strong relationship between the two variables as well.



From the graph we can see that there is an obvious relationship between X and Y and the fitted regression line seems to represent the positive relationship pretty well. However, the data is heteroskedasticity since the variance of residuals is smaller for small values of X and more spread out for large values. Therefore, it violates an important assumption in simple linear regression since we assume the data is homoskedasticity when we initially set up the expression for the appropriate estimators.

	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

	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

Conclusions

By running linear regression with the data set the book provides, we are able to replicate the result and arrive at the conclusion that there exists a strong linear relationship between the TV advertising budget and sales. With statistically significant slope and a reasonable R^2, we can conclude that it is beneficial to the sales when we invest more money on TV advertising.

	TV	Radio	Newspaper	Sales
TV	1.00	0.05	0.06	0.78
Radio	0.05	1.00	0.35	0.58
Newspaper	0.06	0.35	1.00	0.23
Sales	0.78	0.58	0.23	1.00

Table 5: Correlation Matrix

	Statistics	Value
1	Residual Standard Error	1.68
2	R-square	0.90
3	F-Statistics	570.27

Table 6: Regression Quality Statistics