Simple Regression Analysis

Aoyi Shan

Oct 04, 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 namely the scatterplot and essential elements that enable a reproducible workflow to reproduce this simple regression analysis.

Introduction

The main purpose of this project is to find out whether there is a relationship between advertising budget and sales and if so, how strong is the relationship. After we run the linear regression model, we need to 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. We mainly use the TV and Sales columns in this analysis. TV represents the TV Advertising budget in thousands of dollars and Sales is the corresponding product sales in thousands of units in 200 different markets.

Methodology

We start by setting up the linear regression function with TV as the independent variable and Sales as the dependent variable:

$Sales = \beta_0 + \beta_1 * TV$

In order to fit in a line that is as close as possible to the 200 data points we have and minimize the residuals, the Ordinary Least Squares estimator for the slope and intercept should be obtained by running a regression model with the criterion of minimizing the least squares.

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: Regression Coefficients

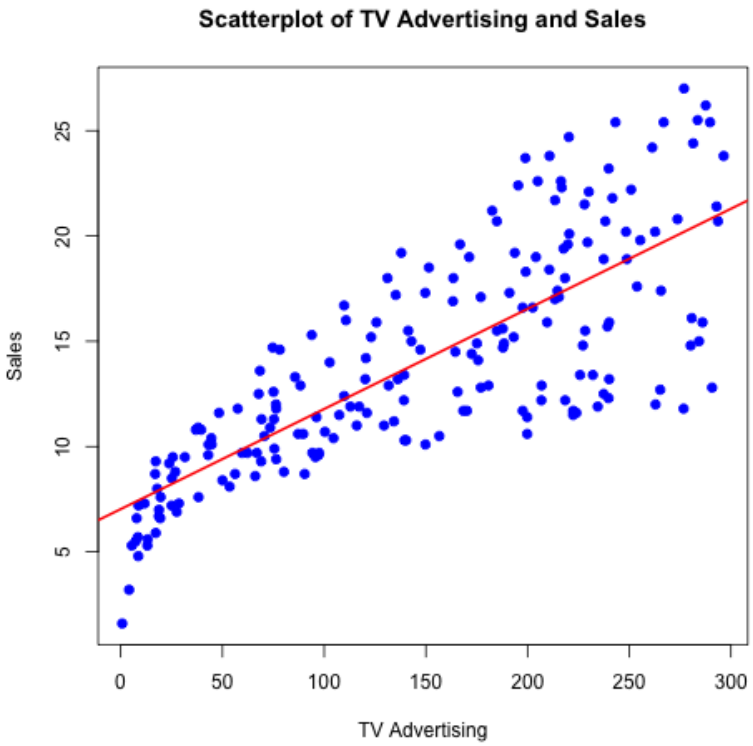
From the table, we can see that the estimated slope is 0.05, which means that for every addition 1000 dollars we spend on TV advertising, on average, we expect to see 47.5 units increase in sales. The intercept can be

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

	Statistics	Value
1	Residual Standard Error	3.26
2	R-square	0.61
3	F-Statistics	312.14

Table 2: Regression Quality Statistics

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.

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.