Simple Regression Analysis

Elly Wang

10/7/2016

Abstract

In section 3.1 - *Simple Linear Regression* of the book **An Introduction to Statistical Learning**, we used the data set for advertising and performed simple linear regression of Sales on TV budget. In this report we will reproduce the graphs, regression, and analysis in a reproducible manner.

Introduction

The purpose of advertising for a certain good is to hopefully to increase its sales. With the technologies available today, there are many possible channels for advertisements to reach their audience. In this report, we will focus on the relationship between the budget alloted for TV advertisement and Sales of a particular good and see if we can determine whether there is a relationship and association between the two variables.

Data

The advertising data used in the book consists of **Sales** (in thousands of units) of a particular prouduct in 200 different markets and advertising budgets (in thousands of dollars) for the product in each of those markets. In particular, the advertising budgets were for **TV**, **Radio**, and **Newspaper**.

Methodology

In this paper, we will be focusing in particular on the relationship between the Sales (in thousands of units) and TV budget (in thousands of dollars). In particular, our model for the simple linear regression will be:

$$\text{Sales} = \beta_0 + \beta_1 * \text{TV}$$

To esimate the coefficients for β_0 and β_1, we will perform the ordinary least squares regression in R.

Results

Using OLS, we get the estimates of the coefficients as shown below in Table 1.

	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: Information about Regression Coefficients

From the table, we can see that for every thousand dollar increase in the budget for TV, sales is predicted to increase by 47.5 units. Since the standard error for the estimate of β_1 is quite low and the t-statistic is quite high, we can say that our estimate is signficiantly different from zero. The estimate of β_0 tells that that even when there are no budget alloted to TV, we'd expect the sales to be at around 7.03 thousands of units.

Regarding the regression quality and the fitness of our regression, we reference the following table (Table 2).

Quantity	Value
Residual standard error	3.26
R Squared	0.61
F-Statistics	312.14

Table 2: Regression Quality Indicies

In this table, we see that the sum of the residuals squared is 3.26, which is a fairly low number for RSS. In doing OLS, we tried to minimize the sum of residuals squared because RSS is directly related to the RSE, which is a measure of lack of fit.
Another way to measure the fitness of a regression is through R^2 From the table, R^2 is 0.61, which means that 61% of the changes in sales is predicted by the changes in TV budget. This number is not considered too high, as 0.99 means close to perfect fit, but it's also not too bad for a fitness test. The high F-statistic in the table also tells us that the estimated coefficents in this regression is significantly different from zero.

The scatterplot of Sales on TV with the regression lines fitted is shown below in Figure 1.

Conclusions

Following the simple linear regression presented in section 3.1 of **An Introduction to Statistical Learning**, we were able to reproduce the graphs, regression model, and arrive at the same results and conclusions.

From the regression, we can see that the linear model we produced using OLS had a fairly good fit. From the siginficance of the regression coefficients and the R^2 value, we can conclude that Sales of the particular good is positively related to the budget alloted for TV advertisment. Although we cannot conclude that its a causal relationship, we can say that these two factors are positively correlated.

Lastly, specifically in the production of this report, we utilized git, github, R, and Makefile to create a streamline workflow that is easily reproducible.

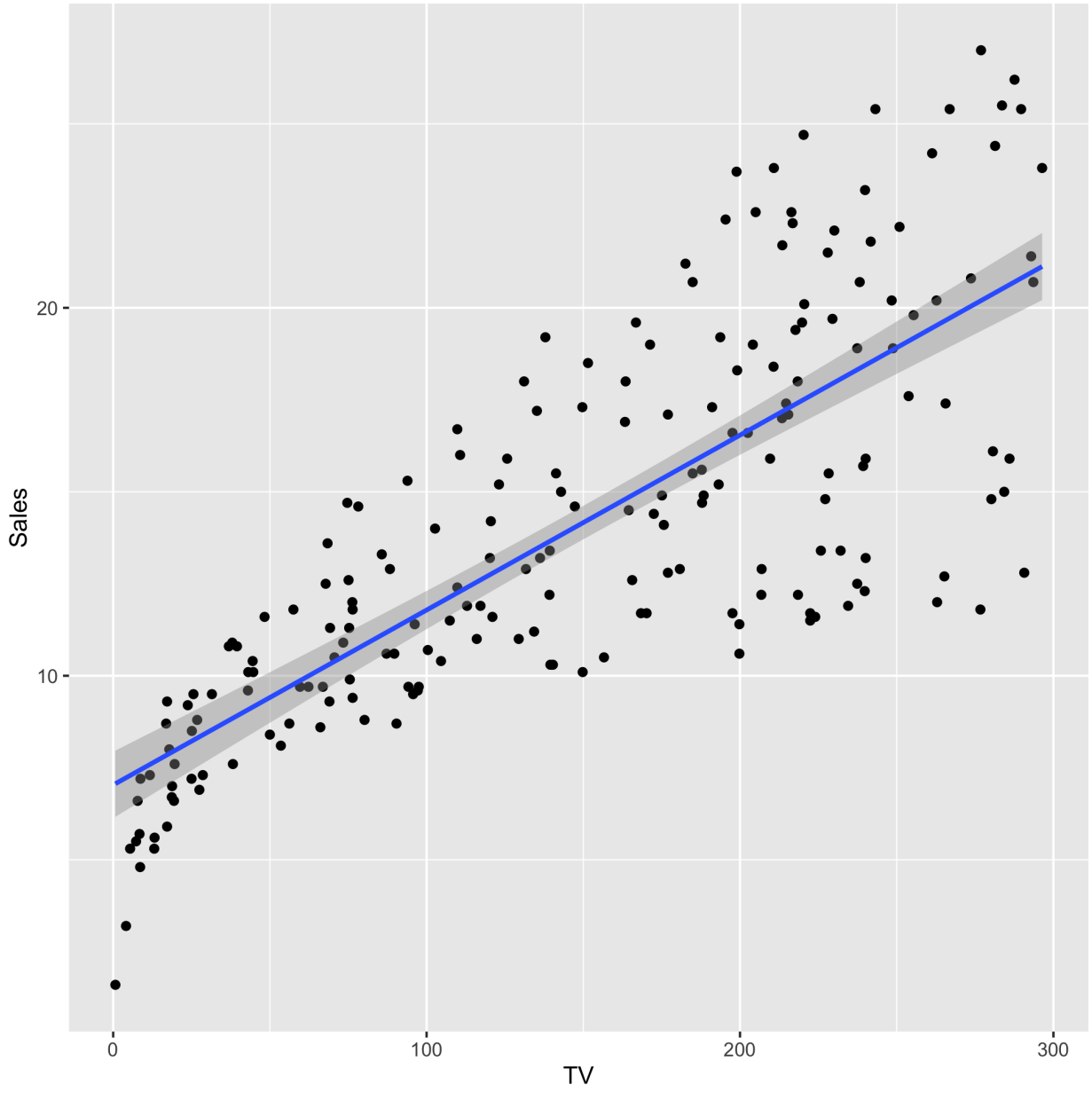

Figure 1: Scatterplot of Sales on TV