Simple Regression Analysis

Erica Wong

October 3, 2016

Abstract

The purpose of this report is to apply the computational tools that we have learned about in class to reproduced a simple regression analysis. Specifically, we are trying to reproduce the analysis that was done in section 3.1 of *An Introduction of Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. You can find a link to their book and data set by clicking on this sentence.

Introduction

The goal of this study is to see how different types of advertising affects sales so that we can provide advice on what medium would be the best to improve sales of a specific product. In my report, I am specifcally looking to see if there is some association between TV advertising budget and sales. If there is an association between the two, then I want to have a model that will be able to predict sales based on the TV advertising budget.

Data

In this report, we are using data from Advertising.csv. In the data set we have four different columns. There is the TV advertising budget, radio advertising budget, and newspaper advertising budget. These budgets are given in thousands of dollars. Additionally, there is a sales column, which is given in thousands of units. In my report, we will be focusing on the TV and sales column of the Advertising.csv.

Methodology

We look at the data set and study the relationship between TV advertising budget and sales. This statement helps us to set up our null and alternate hypothesis. Our null hypothesis is, H_0: There is no relationship between TV advertising budget and sales. Our alternate hypothesis is, H_1: There is a relationship between TV advertising budget and sales. These are equal to the following $H_0 : \beta_1 = 0$ and $H_0 : \beta_1 \neq 0$. To test this, we can use a simple linear regression, where the equation is $Sales = \beta_0 + \beta_1 TV$. In order to solve for β_0 and β_1 we need to fit a linear model to our data, this will give us $\hat{\beta_0}$ and $\hat{\beta_1}$, which are estimates of β_0 and β_1. What this means is that we want to solve for $\hat{\beta_0}$ and $\hat{\beta_1}$ such that we get a line as close as possible to the 200 data points we have. In this report, we will fit a simple linear regression model using the least squares criterion to our data.

Results

Using the least squares criterion, we obtain the following estimates.

Table 1: Information about Regression Coefficents

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.03	0.46	15.36	0.00
TV	0.05	0.00	17.67	0.00

In table 1, we are given a lot of information about our least squares model. Here we can see that $\hat{\beta}_0 = 7.03$ and $\hat{\beta}_1 = 0.05$. Being mindful of the units of each variable, what this means is that for every additional $1000 spent on TV advertising there will be 50 additional units of that product sold. Focusing on $\hat{\beta}_1 = 0.05$, we can see that the standard error for this is rounds to 0, so the standard error is very small. So, the chances that $\beta_1 = 0$ is very small. Also, if we take into account the large t-statistic, then we have enough evidence to reject our null hypothesis. So, there is a relation between TV advertising budget and sales.

We have also provided more information about the least squares model in table 2.

Table 2: Regression Quality Indices

Quantity	Value
RSS	3.26
R2	0.61
F-stat	312.14

RSS is residual sum of squares and it measures the deviation from true regression. Here we can see that our RSS is 3.26. This means that the actual amount of sales will deviate from the true regression by an average of 3260 units. R^2 tells us how much of the dependent variable can be explained by the independent variable. It will always have a value between 0 and 1. In our example, we can see that $R^2 = 0.61$. So, this means that 61% of variability in sales can be explained by TV advertising budget. Finally, there is the F-statistic. The F-statistic can also give us information about our null and alternate hypothesis test. In our case F-stat = 312.14, which is very large given our degrees of freedom. So, we can reject our null hypothesis. The F-statistic supports the same idea that the t-test did in table1. So, all of these things tell us how TV advertising budget relates to sales.

I have also included a picture of the regression line fitted into our scatterplot in Figure 1. From the regression line, we can see that there is a positive correlation between TV advertising budget and sales.

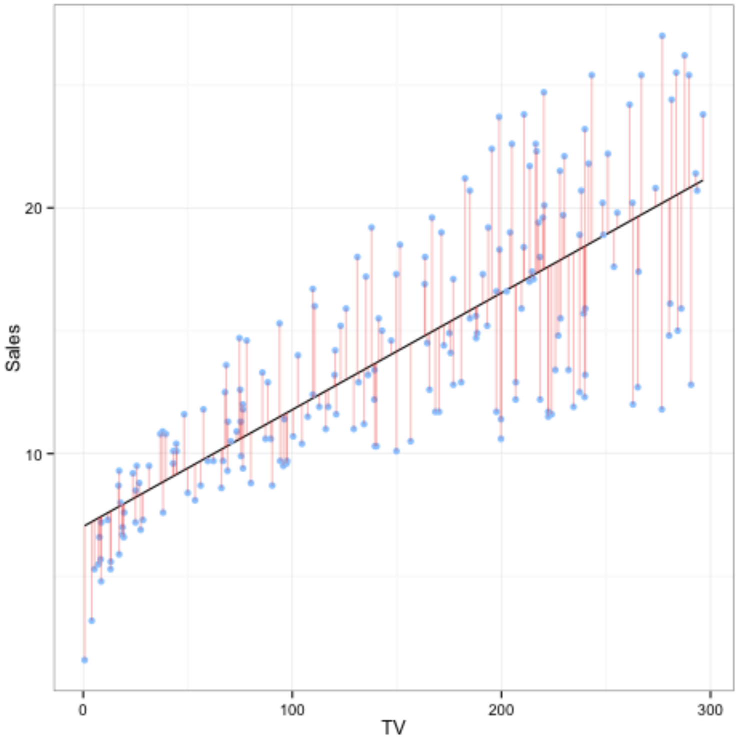

Figure 1: Scatterplot with Fitted Regression Line

Conclusion

So, in this project, we were able to recreate the study that was done in chapter 3 of *An Introduction of Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. We also found that TV advertising budget and sales have a positive correlation between them. This is shown through the positive sloping regression line in Figure 1 and the rejection of the null hypothesis. Our null hypothesis was there is no relationship between TV advertising budget and sales. We know that TV advertising budget and sales have some relationship because R^2 tells us that 61% of the variability in sales can be attributed to TV advertising budget and there are large t and F-statistics, which is what causes us to reject our null hypothesis.

In addition to being able to fit a linear model to the data, we were also able to use the computational tools that we learned about in class to actually work with real data. This homework assignment gave me a better understanding of the different way that these tools actually interact with one another and how powerful each tool actually is. For example, prior to this homework assignment, I did not know that one could run R commands in terminal and Makefile. However, after this project, I know how to write and work with R scripts inside of terminal, which is something that is very cool.