Simple Regression Analysis

Xiaoqian Zhu

Abstract

In this report we reproduce the main results displayed in section 3.1 *Simple Linear Regression*(chapter 3) of the book *An Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Link to the book

Introduction

The overall goal is to provide advice on how to improve sales of the particular product. More specificcaly, the idea is to determine whether there is an association between advertising on TV and sales, and if so, I am going to develop an accurate model that can be used to predict sales on the basis of TV advertising budget by applying a simple linear regression model on advertising data set, compiled by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani.

Data

In this homework, we are working with Advertising data set (Link for Advertising.csv), which consists of *Sales* (in thousands of units) of a particular product in 200 different markets, along with advertising budgets (in thousands of dollars) for the product in each of those markets for three different media: *TV*, *Radio*, and *Newspaper*. In this homework, we are mainly working on the *TV* and *sales* to find out weather they have an association.

Methodology

We consider one media from the data set, *TV*, and the study its relationship with *Sales*. We start our analysis by setting up null and alternative hypothesis. The null hypothesis is, H_0, that there is no relationship between TV advertising budget and sales; the alternative hypothesis is, H_1, that there is a relationship between TV advertising budget and sales. These are equal to the following that $H_0 : \beta_1 = 0$ and $H_0 : \beta_1 \neg 0$. To test the hypothesis, we apply a simple linear regression model, $Sales = \beta_0 + \beta_1 TV$, on the data set, Advertising.csv, to get the estimates of β_0 and β_1, which are $\hat{\beta_0}$ and $\hat{\beta_1}$. In this homework, we will use the least squares model to our data for the regression analysis.

Results

Running regression through R, we can compute get the estimated coefficients. The regression coefficients is given in the table below:

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 can see the estimated coefficients $\hat{\beta_0}$ and $\hat{\beta_1}$. The interception $\hat{\beta_0}$ is equal to 7.03 and the coefficient on TV $\hat{\beta_1}$ is 0.05, which means that for every additional \$1000 spent on TV adverstiing is associated

to 50 additional units of the specific product sold. Morevoer, the t stat of both estimated coefficients are both very high, and the p value are both equal to 0. Thus, both estimated coefficients are statisitcally significant at the 1% level, and we can confidentally reject the null hypothesis. To be noticed, the correlation does not necessarily equal to causation. So we need more inforamtion about the regression analysis to see how much this linear model explain the relationship.

More information about the least squares model from the regression analysis is given in the table below:

Table 2: Regression Quality Indices

	Quantity	Value
1	RSS	3.26
2	R2	0.61
3	F-stat	312.14

RSS is residual sum of squares and it measures the deviation from true regression. In table 2, RSS is 3.26, 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. In table 2, $R^2 = 0.61$, that 61% of variability in sales can be explained by TV advertising budget. F-statistic also describes the statistically difference. In table 2, F-stat is equal to 312.14, which is very hight that we can confidently reject null hypothesis. F-statistic supports the same conclusion from the t-test in table 1. Thus, TV advertising budget and sales have a positive relationship.

I also include a picture of the scatterplot of the analysis with fitted regression line below:

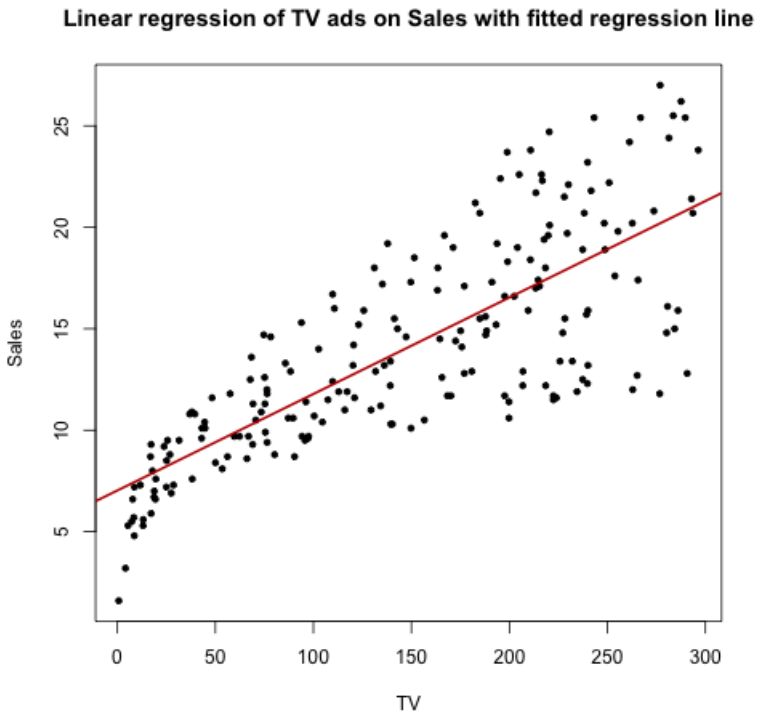

Figure 1: Scatterplot with Fitted Regression Line

From figure 1, we can clearly see a positive correlation between TV advertising budget and sales.

Conclusion

In this homework, we are able to recreate the research of how TV advertising budget affects sales done by in the chapter 3 of chapter 3 of *An Introduction of Statistical Learning*, and derive the conclusion that for every additional $1000 spent on TV adverstiing is correlated to 50 additional units of the specific product sold, by applying a simple linear regression model on the data set, `Adverstiting.csv`. The positive correlation can be easily proved by looking at the positive sloping regression line in Figure 1. The regression analysis gives us both positive and statisitically significant estimated coefficients and can reject the null hyphothesis that TV adversiting budget does not have a correlation with sales. The very high F-statistics also helps to support the same conclusion that TV advertising budget and sales do have a correlation. Additionally, R^2 tells us that 61% of the variability in sales can be explained by this simple linear model, which is TV advertising budget. To be noticed, the correlation does not necessary means causation, it is possible that it is the increase in sales of one specific product lead to the increase in TV advertising budget.

Additionally, this homework also teaches me how to write a scientific report in R. I used to run regression in STATA and copy and paste the tables and figures to Word to compile my work. In this homework, I learned to run codes directly in the report and also can use Makefile to automate my wrapping up process. This saves my lots of time and very cool!