Simple Regression Analysis

Junyu Wang

Oct 7th, 2016

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*.

Introduction

The overall goal is to provide insights about whether advertising through different tunnels improves sales. In this report, we specifically look at how TV advertisement affects sales number. If an association exists between TV advertising and sales, then we want to build an accurate linear model that can be used for sales prediction based on TV advertising budget.

Data

The advertising dataset 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, *Newspaper* and *Radio*). And for this report specifically, we try to discover the association between *Sales* and *Newspaper* in our dataset.

Methodology

We consider *Sales* and TV in our dataset and try to fit them in a simple linear regression model:

$$Sales = \beta_0 + \beta_1 TV$$

And to find the values for the two coefficients β_0 and β_1, we fit the linear regression model based on the normal least square criterion.

Results

We get the regression model's coefficients in Table 1 below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.0326	0.4578	15.36	0.0000
csv_data$TV	0.0475	0.0027	17.67	0.0000

Table 1: Information about Regression Coefficients

From the table above, we can see that the slope of our linear model is around 0.0475 and the intercept point at 7.0326. This indicates that if TV advertisement increases by 1 million dollars, Sales will increase by around 47.5 thousands of units.

But in order to prove that this model is reliable, we need to see if our model is of high quality.

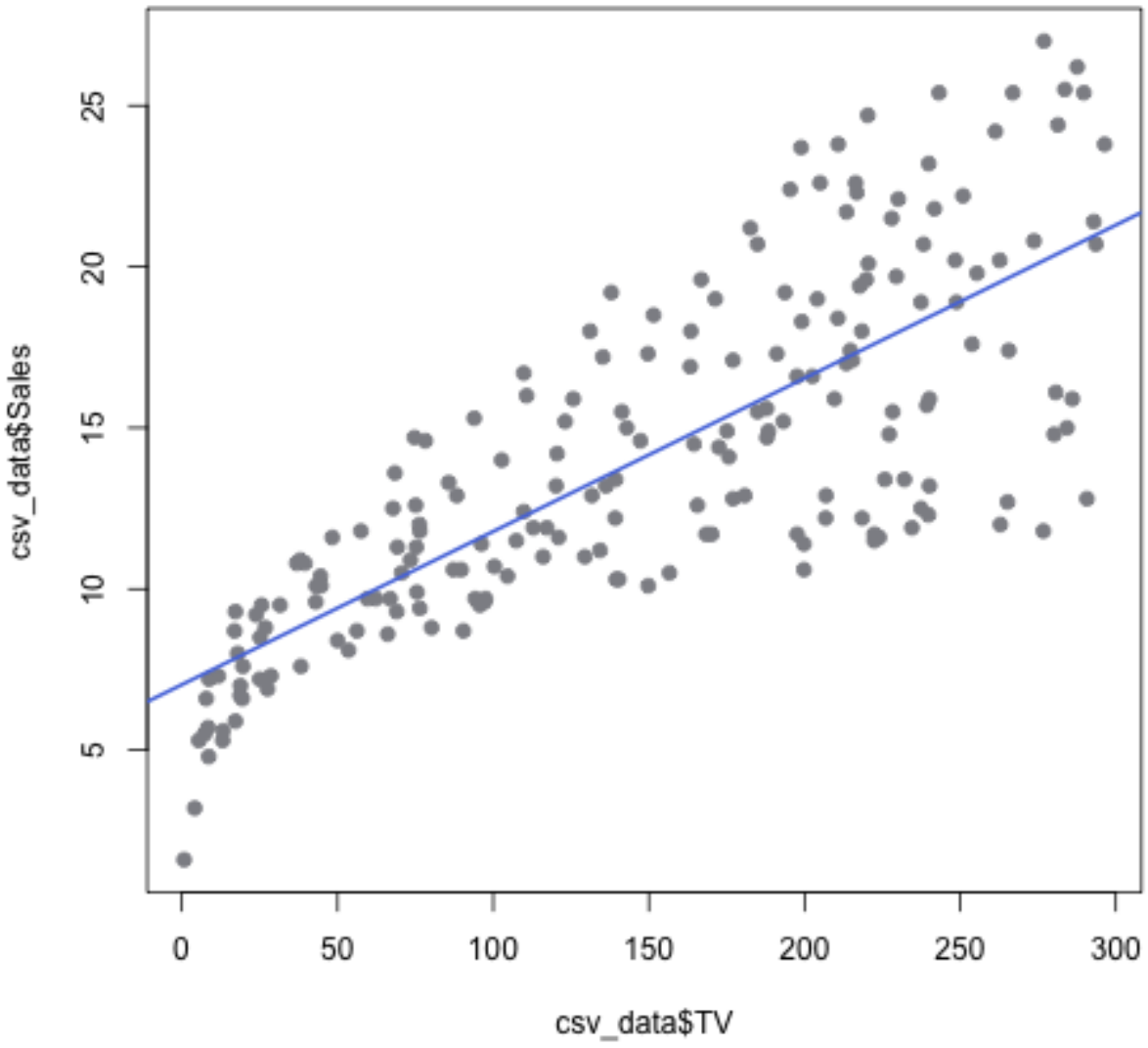
From the table above, we can see that the R squared is 0.61, which is a sign of correlatin between Y and X (in this case *Sale* and TV). Even though this number is not very high, it works for this report because we

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

Table 2: Regression Quality Indicies

don't have many data to train a better linear model. The residual standard error is the average amount that Y deviate from the true regression line, and that means the average deviation of sales from our predicted value is about 3260 units. The mean of Sales is about 14.0 thousand units, and this variation is almost 23.2%. Therefore this shows that the prediction of the trained linear model is not very reliable.

And we can see that the deviation becomes larger as X increases from the scatterplot graph.



Conclusions

Our trained simple linear regression model can make relatively reliable prediction when TV is small and the reliability of prediction decreases as TV gets larger. This problem can be potentially solved by fitting in a larger dataset, remove outliers or use a different kinds of model that fits our dataset better.