Simple Regression Analysis

Priscilla Hartono

October 7th, 2016

Abstract

This paper is the second homework of STAT 159, Reproducible and Collaborative Statistical Data Science, taught by Professor Gaston Sanchez in the fall of 2016 at UC Berkeley.

In this paper we reproduce the main results displayed in section 3.1, *Simple Linear Regression*, of the book **An Introduction to Statistical Learning** (by James et al) using dataset, **Advertising.csv**.

Introduction

Suppose you are a statistical consultant requested to make a marketing plan for next year. You are expected to develop an accurate model that will result in high product sales. What information would be useful in order to provide such recommendations? Would you predict sales on the basis of TV, Radio, and Newspaper? Would you look for a relationship between advertising budget and sales?

Data

The dataset we used in this paper is Advertising.csv. It contains 200 entries of different markets' marketing budget through each media medium (TV, Radio, Newspaper) and how many units were sold.

Sales in thousands of units. Budgets in thousands of dollars

For this paper, we only used the variables TV and Sales. The numbers under TV refers to how much budget was put into advertising by TV, and the number under Sales reflects the number of units sold in that particular market.

Methodology

It turns out that linear regression can be used to develop this model.

Linear Regression

Linear regression is a straightforward approach to predict a linear relationship between X and Y, in this case, TV and Sales. Mathematically, this relationship can be written as

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

Here, β_0 and β_1 are unknown constants or coefficients that represent intercept and slope respectively so we want to get an estimate of $\hat{\beta}_0$ and $\hat{\beta}_1$ for a more accurate measure of future sales.

Estimating Coefficients

The most common approach involves minimizing the least squares criterion. This approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the residual sum of squares (RSS).

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

After some calculus, we can show that the minimizers are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

Assessing the Accuracy of the Coefficient Estimates

Continuing from the above calculation, we want to find the p-value by computing a t-statistic, given by

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

The smaller the p-value, the stronger the association between X and Y and the greater the p-value, the weaker the association between X and Y.

Assessing the Accuracy of the Model

The quality of the linear regression is typically assesed using the residual standard error (RSE) and the R^2 statistic.

RSE is an estimate of the standard deviation of ε, computed by

$$RSE = \sqrt{\frac{1}{n-2}RSS}$$

It provides an absolute measure of lack of fit of the model to the data. This is measured in the units of Y, hence, it is not always clear what a good RSE constitutes. Meanwhile, R^2 statistic takes the form of a proportion, so it always takes on a value between 0 and 1, and is independent of the scale of Y.

$$R^2 = \frac{TSS - RSS}{TSS}$$

where TSS is the total sum of squares.

An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression, and an R^2 statistic that is close to 0 indicates that the regression did not explain much of the variability in the response.

Results

In Table 1, we computed the regression coefficients and in Table 2, more information about the least squares model can be found.

We can look at the p-value and determine whether there is a relationship between X and Y. In this case, since the p-value is close to 0, we can conclude that there is a correlation between TV and Sales. The RSE

Table 1: Information about Regression Coefficients

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

Table 2: Regression Quality Indices

	Quality	Value
1	RSE	3.26
2	R2	0.61
3	F-statistic	312.14

value is to determine on average how much the actual sales deviate from the true regression. Here, RSE is 3.26, meaning, on average, the actual sales deviates from the true regression by 3.26 units.

In Figure 1, the scatterplot can be found. We know that if there is a positive slope, there is a positive correlation and if there is a negative slope, there is a negative correlation. From this data, we can see a positive slope, showing that there is a positive correlation between TV and Sales.

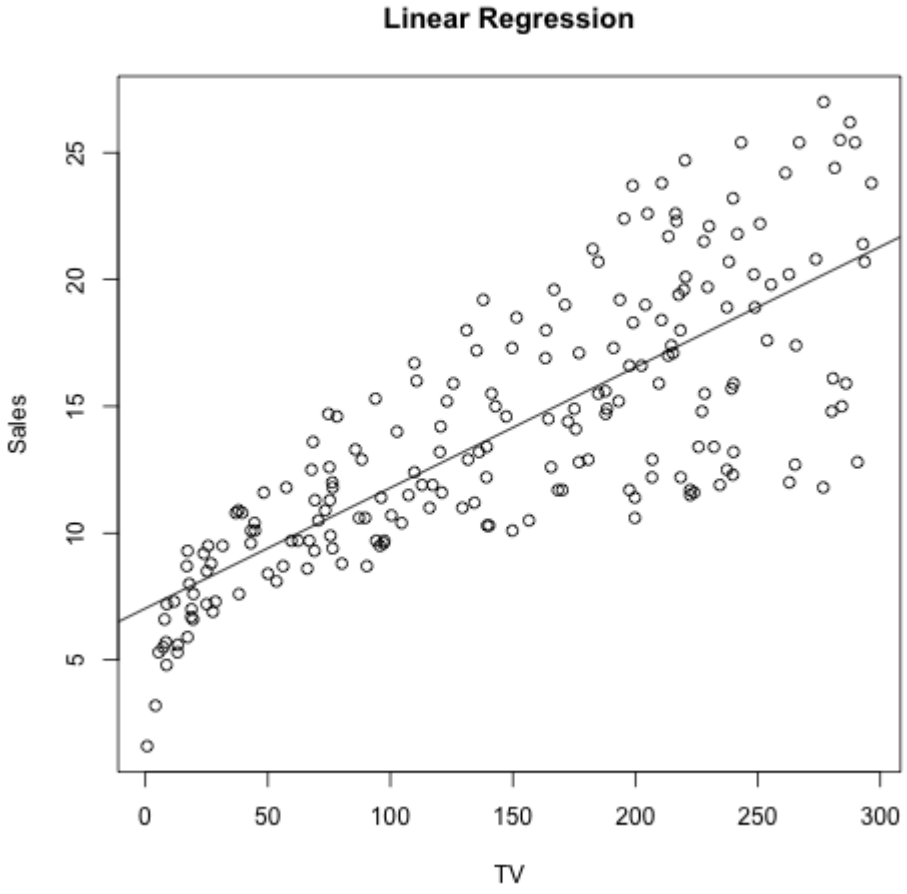

Figure 1: Scatterplot with fitted regression line

Conclusion

In conclusion, linear regression model can be used to determine whether or not there is a relationship between 2 variables. With this information, a marketing model could be produced, in hopes of resulting in high product sales.

With this particular data in mind, we can conclude that there is a positive correlation between TV and Sales. For every $1000 increase in TV advertising, approximately 47.5 units of product is sold.