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. We want to use our data to produce estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ for a more accurate measure of future sales, where $\hat{y}$ indicates a prediction of Y on the basis of X = x.

Estimating Coefficients

Using data we have, we want to estimate the coefficients. The most common approach involves minimizing the least squares criterion. Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X and $e_i = y_i - \hat{y}_i$ represents the difference between the ith observed response value and the ith response value that is predicted by our linear model.

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

This approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the residual sum of squares (RSS). 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 t-statistic, given by

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

t-statistic measures the number of standard deviations that $\hat{\beta}_1$ is away from 0. Consequently, the probability of the p-value can be calculated by observing any value greater than or equal to |t|. 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. When the null hypothesis is rejected, it means that an association between X and Y exists, since the p-value is small enough, about 1%.

Assessing the Accuracy of the Model

Once we have rejected the null hypothesis, we want to quantify the extent to which the model fits the data. The quality of the linear regression is typically assesed by the residual standard error (RSE) and the R^2 statistic.

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

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

where

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

It roughly gives the average amount that the response will deviate from the true regression line. 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, 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 $TSS = \sum(y_i - \hat{y}_i)^2$ is the total sum of squares.

RSS measures the amount of variability that is left unexplained after performing the regression. Hence, TSS - RSS measures the amount of variability in the response that is explained by performing the regression, and R^2 measures the proportion of variability in Y that can be explained using X.

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.

The R^2 statistic is a measure of the linear relationship between X and Y, similar to correlation. This suggests that we might be able to use r = Corr(X, Y) instead of R^2 to fit the linear model. In fact, we can use $R^2 = r^2$

Results

From the formulas above, we can calculate the regression coefficients and least squares model to determine whether or not there is a correlation between X and Y (TV and Sales) and by how much.

Table 1 computes the regression coefficients. Looking at the p-value, we can 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.

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 computes the least squares model. The RSE value determines on average, how much the actual sales deviates from the true regression. Here, RSE is 3.26, meaning, on average, the actual sales deviates from the true regression by 3.26 units.

Table 2: Regression Quality Indices

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

If the tables above shows that there is a correlation between X and Y, we want to analyze Figure 1 to determine if the correlation is positive or negative. Figure 1 outputs a scatterplot. 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 between TV and Sales, hence, showing that there is a positive correlation between them.

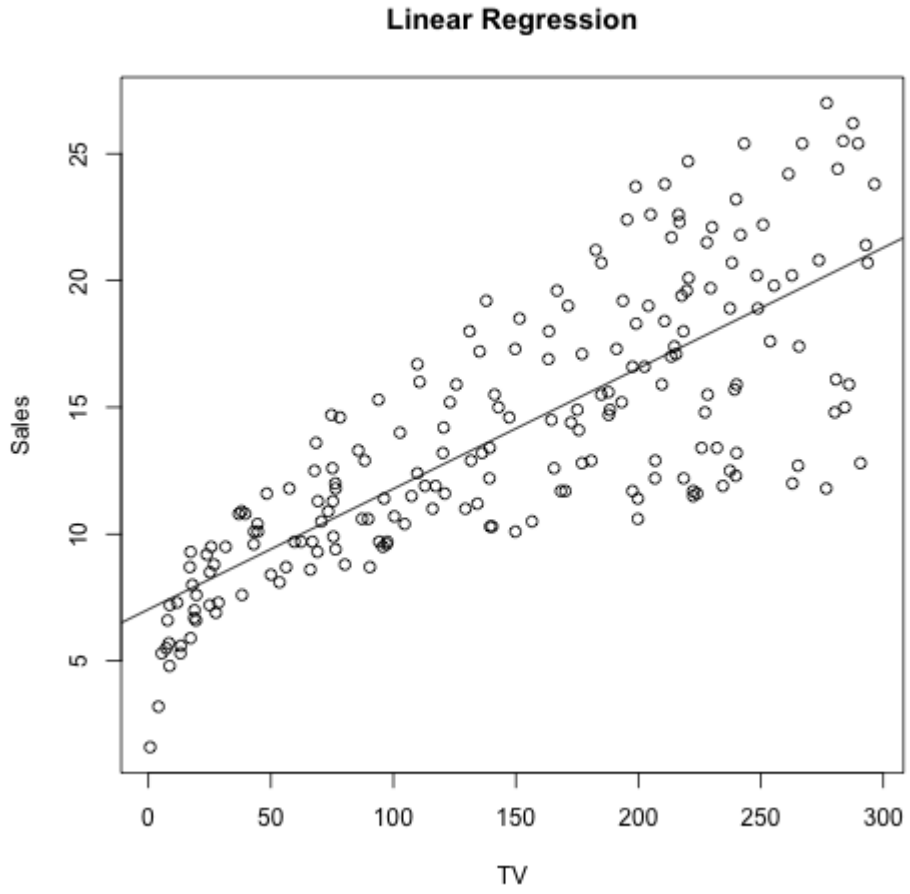

Figure 1: Scatterplot with fitted regression line

Conclusion

By observing the p-value of this particular data, we can conclude that there is a correlation between TV and Sales, and it is a positive one. From Table 1, we can also point out that for every $1000 increase in TV advertising, approximately 47.5 units of product is sold.

As a general idea, linear regression model can be used to determine whether or not there is a correlation between 2 variables, and if so, by how much. With this information, a marketing model can be produced, in hopes of resulting in high product sales.