Simple Regression Analysis

Stacy J Chang

October 7th, 2016

Abstract

Fall 2016 Statistics 159, Reproducible and Collaborative Statistical Data Science, homework two introduced simple regrssion analysis. In this report, I utilized the Advertising CSV file that was provided to reproduce the results displayed in chapter 3, *Simple Linear Regression*, of the text book, **An Introduction to Statistical Learning**.

Introduction

The overarching goal of this homework is for the students to get familiarized with linear regression, a simple yet powerful tool to analysize data. Even though there is multiple regression analysis, this report primaryly uses simple linear regression, which only considers two variables. The report focuses on the potential relationship or regression between **Sales** and **Advertisment** in three different media outlets.

Data

The Advertising data that used in this report was provided through this link, which is part of the textbook, **An Introduction to Statistical Learning** written by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Here is a small section of the Advertising data:

```
##   X    TV Radio Newspaper Sales
## 1 1 230.1  37.8      69.2  22.1
## 2 2  44.5  39.3      45.1  10.4
## 3 3  17.2  45.9      69.3   9.3
## 4 4 151.5  41.3      58.5  18.5
## 5 5 180.8  10.8      58.4  12.9
## 6 6   8.7  48.9      75.0   7.2
```

The data represent **Sales** through three different media: **TV, radio, and newspaper**, in 200 different markets. The data that the paper focuses on are the **Sales** and **TV** data points. **Sales** data represents the total amount of profit, in thousands of units. While **TV** reflects the cost, in thousands of units, in total spent on TV advertisement.

Methodology

Simple linear regression is a very simple approach for estimating a quantitative response Y on an independent variable X. The regression assumes that there is a linear relationship between Y and X. The regression is usually model after this simple equation:

$$Y_i \approx \beta_0 + \beta_1 X_i \tag{1}$$

Estimating the Coefficients

In this paper, we concentrated on one particular media platform, which is TV, and its relationship with **Sales**. The linear model that the we perform the simple linear regression on is:

$$Sales \approx \hat{\beta}_0 + \hat{\beta}_1 TV \tag{2}$$

In order to accurately estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ that fit our linear model, we needed to use the 200 samples that are given in the data. Ultimately, our goal is to fine the *closeness* between $\hat{\beta}_0$ and $\hat{\beta}_1$ and $n = 200$ different markets. We have to find an intercept of $\hat{\beta}_0$ and a slope $\hat{\beta}_1$ such that the resulting line is as close to the $n = 200$ data points.

We used *residual sum of squares* (RSS) to estimate the least squares fit for the regression. Let $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ be the prediction of Y, and the _i_th *residuals* is $e_i = y_i - y_i$. *Residual sum of squares* (RSS) is:

$$RSS = e_1^2 + e_2^2 + \cdots + e_3^2 \tag{3}$$

Utilizing this knowledge, we found that $\hat{\beta}_0 = 7.03$ and $\hat{\beta}_1 = 0.0.475$. These values translate to that with an additional of $1000 spent on TV advertisement is associated with selling approximately 47.5 additional units of product.

Assesing the Accuracy of the Coefficients Estimates

The observed β values can be used to come up with a hypothesis test on whether or not there is a relationship between X and Y. Our null hypothesis would be that there is no relationship between the two variables and we perform a *t-test* in order to test the hypothesis:

$$t = \hat{\beta}_1 - 0/SE(\beta_1) \tag{4}$$

The result of the above equation is the p-value. With a small p-value, typically less than 0.05, we would be able to reject our nulll hypothesis and conclude that there is a relationship between the two variables.

RSE

RSE or *residual standard error* estimates the standard error or the given model. The equation for it is:

$$\sqrt{}(1/(n-2) * \sum_{i=1}^{n} (y_i - yhat_i)^2) \tag{5}$$

R^2

R^2 value measures how close the data are to the fitted regression line. The value of R62 lies between zero and one. equation of R^2 is:

$$R^2 = (TSS - RSS)/TSS \tag{6}$$

Results

The result that we obtain for $\hat{\beta}_0$ and $\hat{\beta}_1$ is described in the below table:

Table 1: Information about Regression Coefficients

	Estimate	Std. Error	t value	P-value
(Intercept)	7.03	0.46	15.36	< 0.0001
TV	0.05	0.00	17.67	< 0.0001

As one can see, the p-values are all less than 0.0001, which means that we can reject the null hypothesis, in other words, there is a relationship between X and Y.

The table below describes additional information about the least squares model for the regression:

Table 2: Regression Quality Indices

Quantity	Value
RSE	3.26
R2	0.61
F-stat	312.14

The RSE value is about 3.26, which means that on average actual sales in each market deviate from the true regression line by approximately 3,260 units.

Below is the scatterplot showing the relationship between **Sales** and **TV**:

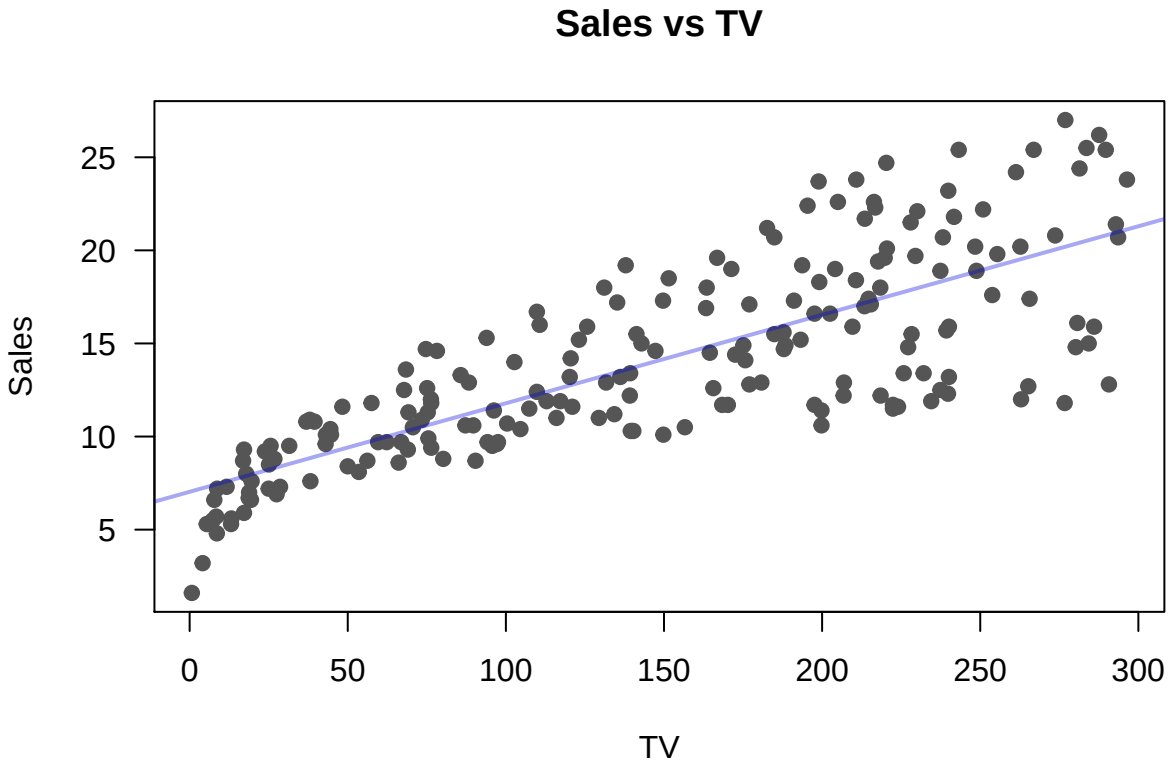

Figure 1: Scatterplot of **Sales** and **TV** with fitted line

As one can see the positive slope of the regression line placed in the center of the data points, which means there is a positive correlation with the two variables.

Conclusions

In conclusion, using simple linear regression model we found the relationship between **Sales** and **TV** advertisement. We conclude that approximately with an increase of $1000 spendingi in TV advertisment is equalvilant as selling approximately 47.5 units of product. With an relatively easy linear regression model, we were able to predict and estimate powerful data that are essential to successfully analysize the data. The linear regression relationship between X and Y variables is the key aspect to the paper.