Simple Regression Analysis

Andrea Widjaja

10/07/16

Stat 159 Homework 2

Abstract

This paper consists of the reproduced results from *Section 3.1 Simple Linear Regression* from the book **An Introduction to Statistical Learning** *by Gareth James, Daniel Witten, Trevor Hastie, and Robert Tibshirani.* This paper will discuss topics of simple linear regression such as estimating the coefficients, assessing the accuracy of the coefficient estimates, and assessing the accuracy of the model using the least square approach.

Introduction

Suppose we are given information regarding the sales progress of a particular product (in thousands of units) and advertising budgets (in thousands of dollar) for three medias. As statistical consultants, we are given a task to develop a marketing plan that will result in high product sales with minimum advertising budget possible. Before tackling this problem, there are several factors that we must consider. First of all, we should determine whether there is a relationship between advertising budget and sales. If there is no relationship, then there is no point in spending money on advertising. If there is, however, a relationship, then we should know the strength of this relationship. Meaning that, if we are given the advertising budget, are we able to predict sales with a high level of accuracy?

This is when Simple Linear Regression steps in. Linear regression can be utilized for predicting a quantitative estimate to provied a data-based solution.

Data

The data set we are given is *Advertising.csv*, downloaded from `http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv`. This data set contains the sales(in thousands of units) of a product across 200 different markets and the amount spent(in thousands of dollars) in advertising for the product in each market. The advertising data contains budgets for three different medias: TV, radio and newspaper. It has 200 records and 5 variables(index, TV, Radio, Newspaper, Sales). Each record contains the information of advertising budget for TV, Radio, and Newspaper, including the sales progress for each market.

For this paper, we will focus on only the TV media and analyze its relationship to sales.

Methodology

Linear Regression

Suppose that we want to estimate a relationship between two variables X and Y. We assume that there is a linear relationship between the Variable X and Variable Y. For example, if we presume that a change in X

has an effect on Y, then we can say that X is the independent variable, whereas Y is the dependent variable. This relationship can be represented as a simple lienar model.

Regress Y on X:

$$Y = \beta_0 + \beta_1 X$$

This, however, do not take into account for the error term/residual. The dependent variable Y, may not be perfectly predicted by the variable X. The error term/residual represents the difference between the results obtained by the model and real-world applications.

If we take into account for error, the linear regression will be:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

For our case, we need to analyze the relationship between TV advertising and Sales. The first step we need to do is come up with a similar linear model. X, for example, can represent TV, whereas, Y can represent Sales.

Regress Sales on TV advertising, taking into account for residual:

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

Estimating the Coefficient

From the linear models above, β_0 and β_1 are constants. β_1 determines the slope, and β_0 determines the intercept. They are also known as coefficients or parameters.

Unfortunately, in practice, these betas are unknown. We must use data to estimate them. Let's say that we observed n pairs of data:

$$(TV_1, Sales_1), (TV_2, Sales_2), ..., (TV_n, Sales_n)$$

We have to find the coefficient intercept estimate $\hat{\beta}_0$ and slope estimate $\hat{\beta}_1$ such that the linear regression fits the data well. To achieve the closest line possible to the data points, we can use the least square criterion.

Let the predicition for Y based on ith value of X be:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * x_i$$

and ith residual (estimate of the error term) be:

$$\epsilon_i = y_i - \hat{y}_i$$

Residual Sum of Squares (RSS):

$$RSS = \sum_{i=1}^{n} (\epsilon_i)^2 = \epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2 + ... + \epsilon_n^2$$

or simply,

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + ... + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

For the least square approach, solving for $(\hat{\beta}_0, \hat{\beta}_1)$ minimizes the RSS.

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

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1} * \bar{x}$$

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

We get the coefficient (in **Table 1**): $\hat{\beta_0} = 7.03$ and $\hat{\beta_1} = 0.0475$

This solves for the least squares coefficient estimates for the simple linear regression.

Hypothesis Tests

Asessing the Accuracy of the Coefficient Estimates

Since standard errors can measure accuracy, standard errors are used to perform *hypothesis tests* on the coefficients. We are going to test the *null hypothesis* versus the *alternative hypotehsis*.

Null hypothesis :

$$H_0 : ThereIsNoRelationshipBetweenXandY$$

$$H_0 : \beta_1 = 0$$

versus

Alternative hypothesis:

$$H_a : ThereIsSomeRelationshipBetweenXandY$$

$$H_a : \beta_1 \neq 0$$

When testing the null hypothesis, we need to know how far $\hat{\beta_1}$ is from 0. This depends on the standard error of $\hat{\beta_1}$. If the $SE(\hat{\beta_1})$ is small, then we can conclude that $\beta_1 \neq 0$ and that there is a relationship between X and Y in the linear model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

.

On the contrary, if $SE(\hat{\beta_1})$ is large, then β_1 is large, such that we have to reject the null hypothesis.

t-statistic:

$$t = \hat{\beta_1} - 0/SE(\beta_1)$$

measures the number of standard deviations that $\hat{\beta_1}$ is away from 0. If there is no relationship betwen X and Y, then the equation will be

$$Y = \beta_0 + \epsilon$$

Thus, we can say that $\hat{\beta_1} = 0$, and that we will have a t-distribution with $n - 2$ degrees of freedom. We can then find the $p - value$, which is the probability of observing any value equal to $|t|$ or larger. If the p-value is small, then there is a relationship between X and Y, and we reject the *null hypothesis*.

(results can be seen in **Table 1**)

Assessing the Accuracy of the Model

If we reject the null hypothesis, and go for the alternative hypothesis instead, we should measure the extent to which the model fits the data. The quality of a linear regression fit is assessed using two related quantities: *Residual Standard Error* and $R^2 Statistic$.

Residual Standard Error(RSE)

The RSE is an estimate of the standard deviation of ϵ. It is the average amount that the response will deviate from the true regression line. Additionally, in simple words, it is a measure of the lack of fit of the linear model to the data.

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

$$RSE = \sqrt{(\frac{1}{n-2}) * (\sum_{i=1}^{n}(y_i - \hat{y}_i)^2)}$$

When the RSE is small, then we can assume that for $i = 1, ..., n$, $\hat{y}_i = y_i$. This means that we can conclude that the model fits the data very well. On the contrary, if $\hat{y}_i$ is very far from y_i, for one or more of the observations, then the RSE will be large, indicating that the model does not fit the data very well.

R^2 Statistic

The $R^2 Statistic$ provides an alternative measure of good fit to the RSE. It is a measure of the linear relationship between X and Y, and has the range [0,1]. Additionally, it is identical to the squared correlation. The formula of R^2 is given by:

$$R^2 = \frac{(TSS - RSS)}{TSS} = 1 - \frac{RSS}{TSS}$$

where,

$$TSS = \sum(y_i - \hat{y}_i)^2$$

is the total sum of squares.

If R^2 is near 0, then our linear model does not fit the data well. On the contrary, if R^2 is near 1, the our linear model fits the data well.

(results can be seen in **Table 2**)

Results

This scatterplot (**Figure 1**) shows the least squares fit for the regression of Sales onto TV. The line represents a simple model that can be used to predict *Sales* using the TV medium.

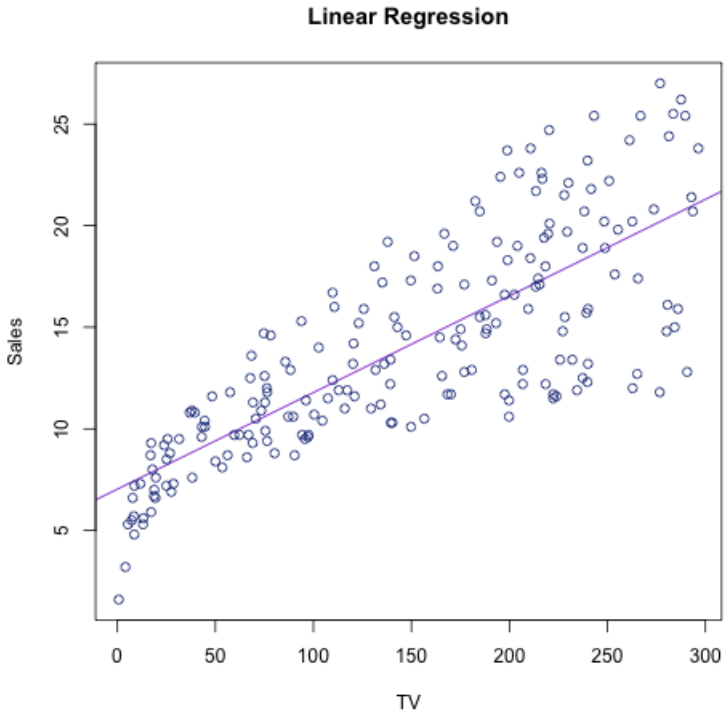

Figure 1: Scatterplot with Fitted Regression Line - From the Advertising data, this shows the least squares fit for the regression of Sales onto TV. This means that, the purple line represents a simple model that can be used to predict sales using TV.

To access the accuracy of the coefficient estimates, several calculations are needed to be made to solve for $\hat{\beta}_0$ and $\hat{\beta}_1$. We achieved the regression coefficients in the following table (**Table 1**):

Table 1: Information about Regression Coefficients - note that: Sales variable is measured in thousands of units, and the TV variable is in thousands of dollars

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.0326	0.4578	15.3603	0.0000
TV	0.0475	0.0027	17.6676	0.0000

The first column of Table 1 shows the coefficients we estimated. The estimated slope coefficient $\hat{\beta}_1$ is 0.047, since it is measured in thousands of units, then for every $1,000 of increase in budget sales, there will be an increase in total sales by $48. With the p-value being very low (almost 0), this shows that TV advertising effect Sales greatly.

To access the accuracy of the model, there are several calculations that are needed to be made. More information about the least square model for the regression of number of units sold on TV advertising budget is given in the table (**Table 2**) below:

Table 2: Regression Quality Indices

	Quantity	Value
1	RSE	3.2587
2	R^2	0.6119
3	F-statistic	312.1450

The R Squared is 0.61, meaning that 61% of Sales is affected by a linear regression on TV. Since R^2 is closer to 1, it means that our linear model fits the data well. Additionally, with the value of RSE being 3.26, it shows that the sales in each market deviates from the regression line approximately 3,260 units on average. Since RSE is a measure of bad fit, and it is small, it shows that our linear model fits the data pretty well.

Conclusions

In conclusion, it has been proven that linear regression can be utilized to predict the outcome of an independent variable from a dependent variable, if there happens to be a relationship between the two variables. Additionally, we can see that from Table 1, that the intercept $\hat{\beta}_0 = 7.03$ and the slope $\hat{\beta}_1 = 0.0475$. This means that an increase of \$1,000 in the TV advertising budget, is associated with an increase in sales by around 50 units. To see whether or not this is a good number depends on the client.