ISSN:0975-9646

Simple Linear Regression with Least Square Estimation: An Overview

Aditya N More[#1], Puneet S Kohli[*2], Kshitija H Kulkarni[#3]

[#1-2]*Information Technology Department,*[#3] *Electronics and Communication Department*
College of Engineering Pune
Shivajinagar, Pune – 411005, Maharashtra, India

***Abstract*— Linear Regression involves modelling a relationship amongst dependent and independent variables in the form of a linear equation. Least Square Estimation is a method to determine the constants in a Linear model in the most accurate way without much complexity of solving. Metrics such as Coefficient of Determination and Mean Square Error determine how good the estimation is. Statistical Packages such as R and Microsoft Excel have built in tools to perform Least Square Estimation over a given data set.**

***Keywords*— Linear Regression, Machine Learning, Least Squares Estimation, R programming**

I. INTRODUCTION

Linear Regression involves establishing linear relationships between dependent and independent variables. Such a relationship is portrayed in the form of an equation also known as the linear model. A simple linear model is the one which involves only one dependent and one independent variable. Regression Models are usually denoted in Matrix Notations. However, for a simple univariate linear model, it can be denoted by the regression equation

$$\hat{y} = \beta_0 + \beta_1 x + \varepsilon$$

where
y^ is the dependent or the response variable
x is the independent or the input variable
β_0 is the value of y when x=0 or the y intercept
β_1 is the value of slope of the line
ε is the error or the noise

This linear equation represents a line also known as the 'regression line'. The least square estimation technique is one of the basic techniques used to guess the values of the parameters β_0 and β_1 based on a sample set.

II. LEAST SQUARES ESTIMATION

This technique estimates parameters β_0 and β_1 by trying to minimize the square of errors at all the points in the sample set. The error is the deviation of the actual sample data point from the regression line. The technique can be represented by the equation.

$$\min \sum_{i=0}^{n} (y_i - \hat{y}_i)^2 \qquad (2.1)$$

where

y_i is the i[th] value of the sample data point

$\hat{y}$ is the i[th] value of y on the predicted regression line

The above equation can be geometrically depicted by figure 2.1. If we draw a square at each point whose length is equal to the absolute difference between the sample data point and the predicted value as shown, each of the square would then represent the residual error in placing the regression line. The aim of the least square method would be to place the regression line so as to minimize the sum of the areas of all such squares.

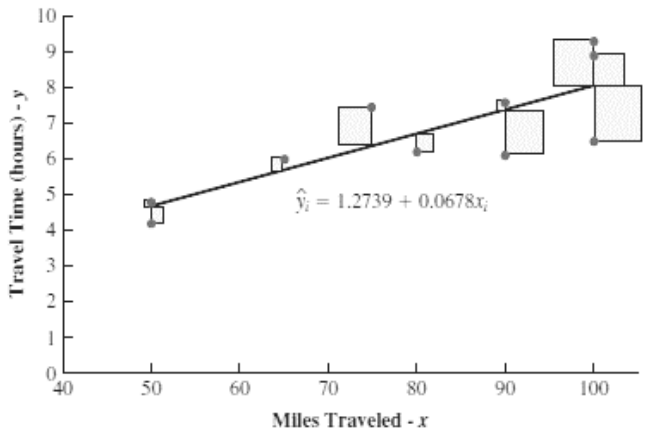

Fig. 2.1 Least Square Estimation can be picturized as an attempt to reduce the area of the squares whose length is equal to the y axis deviation of the point

Using differential calculus on equation *2.1* we can find the values of β_0 and β_1 such that the sum of squares is minimum.

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (2.2)$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \qquad (2.3)$$

where

$\bar{x}$ is the mean value of x

$\bar{y}$ is the mean value of y

Once the Linear Model is estimated using equations *2.3* and *2.4*, we can estimate the value of the dependent variable y in the given range only. Going outside the range is called extrapolation which is inaccurate if simple regression techniques are used.

III. IMMEDIATE CALCULATION IN LEAST SQUARE ESTIMATIONS

The calculations for least square estimation involves immediate values called the 'Sum of Squares'[1] which can help us understand how well does the linear model summarize the relationship between the dependent and independent variable.

A. SSE

The sum of squares due to errors denotes the error in estimating the values of the dependent variable in the sample. This is the value we try to minimize during in Least Squares estimation. It can be expressed as

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2 \quad (3.1)$$

In short, SSE explains how well do the points cluster around the regression line.

B. SST

If we consider the mean of all the dependent variable values in the sample, we can find out how much does every sample value of the dependent variable deviate from the mean value. The Total Sum of Squares is such a measure and can be expressed as

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad (3.2)$$

C. SSR

We can find out how much do the points on the regression line deviate from the mean value using the sum of squares due to regression. It is expressed as

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad (3.3)$$

D. COEFFICIENT OF DETERMINATION

From equations 3.1, 3.2, 3.3 we can observe that all these values are related and it can be explicitly stated that

$$ST = SSE + SSR$$

Thus the sum of squares due to regression and error add up to the total sum of squares. This shows that if any two of these squares is known, the third can be calculated easily.

The goodness of fit for the linear model can be determined by the variable known as 'The coefficient of determination' [2] which can be expressed as

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (3.4)$$

The value of r^2 can vary from 0 to 1 and the more the value, the better fit is the linear model. In practical cases, a coefficient of determination of 0.25 is also considered acceptable.

E. STANDARD DEVIATION ABOUT THE REGRESSION LINE

Deviation around the regression Line can be expressed by the Mean Square Error which is the average square of error around the regression line.

$$MSE = \frac{SSE}{n-2} \quad (3.5)$$

IV. SIMPLE LINEAR REGRESSION USING TOOLS

A. R PROGRAMMING

R supports linear regression over a given data set through various built in commands. Once we import all the data in R, we can run the lm() command on the data to obtain the linear model. [3]

Ex.

```
> model = lm(Time ~ Distance)
> summary(model)
```

This returns Residuals and Coefficients where we can obtain the values for β_1 and β_2. The fitted values can be obtained by using the fitted() command and the residuals can be obtained by using resid() command.

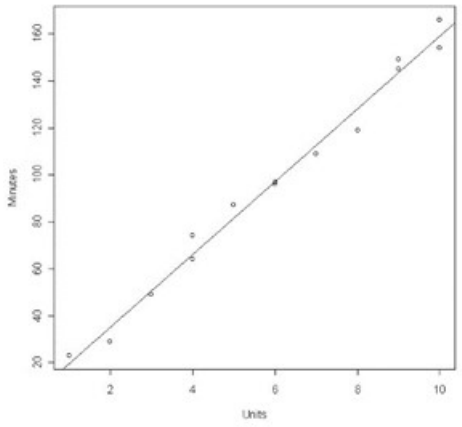

Fig. 4.1 Regression Line plotted using abline()

The regression line can be then plotted as shown in figure 4.1 simply by using the command

```
> abline(model)
```

This line can also be adjusted to make it pass through the origin if required by just saying

```
model = lm(Time ~ 0 + Distance)
```

B. *MICROSOFT EXCEL*

Microsoft Excel provides Chart Tools which can be used to compute the Regression Equation and calculate parameters like the coefficient of determination. The following steps can be followed to determine the Estimated Regression Equation:

1. Right Click on the data chart and select 'Add Trendline'
2. In the 'Format Trendline' Taskpane
   a. In the 'Trendline Options' area, select 'Linear'
   b. Select 'Display Equation on chart'

To view the Coefficient of determination

3. In the 'Trendline Options' area, select 'Diplay R-squared value on chart'

V. ADVANTAGES AND LIMITATIONS OF LEAST SQUARE ESTIMATION

Linear Least square estimation is usually very optimal in nature and can help obtain good results in a very limited data set. Also, there are many linear estimation methods which involve considering the absolute difference between the regression line and sample data point. However, such differences may cancel each other out. If we attempt taking absolute differences, the complexity of differentiation increases. Compared to these methods, Least Square Estimation proves to be a simpler and more accurate solution.

However, as discussed earlier, Least Square Estimation does not work well for extrapolation. This technique being very sensitive to the estimation errors, the regression line may change drastically as the sample points increase. Hence beyond the sample space, we have no guarantee that the regression model still holds. Also, this technique gets easily affected by outliers. Even one or two outliers may change the placement of the regression line drastically.

VI. CONCLUSION

The simple Least Squares Estimation for univariate Regression discussed above is not sufficient to be used in practical scenarios where there are multiple independent variables involved. However multiple regression techniques are based on the same principles as that of a simple regression technique. Matrices are heavily used in such scenarios. Also in certain scenarios, a multiple regression model is converted to a simple model by removing the effects of the other variables.

Though the least squares estimation is heavily affected by outliers and cannot be sufficiently used to extrapolate data, it is still popularly used to estimate linear models. Linear Regression continues to serve many applications in the fields of social science, finance, biology etc.

REFERENCES

[1] Camm, Cochran, Fry, Ohlmann, Anderson, Sweeney, Williams, Essentials of Business Analytics, 1st Edition, Cengage Learning.J.

[2] George A. F. Seber, Alan J. Lee, Linear Regression Analysis, 2nd Edition, January 2003, Wiley.

[3] John O. Rawlings, Sastry G. Pantula, David A. Dickey, Applied Regression Analysis : A Research Tool, May 2001, Springer.