Detailed Information on SLRM and MLRM

INTRODUCTION

SLRM & MLRM are the two most beautiful and important topics of Statistical Inferences. Here in this reading, we first go with some basic concepts and Terminologies that we have come across in the Readme file of this project. This will help the reader by understand the project even if the reader has no past knowledge of Regression Models.

HISTORY OF LINEAR REGRESSION MODELS

Linear regression is a statistical method that uses a linear equation to model the relationship between a dependent variable and one or more independent variables. The history of linear regression can be traced back to the work of Adrien-Marie Legendre and Carl Friedrich Gauss in the early 19th century. Legendre developed the method of least squares, which is a way to fit a line to a set of data points that minimizes the sum of the squares of the residuals. Gauss independently developed the same method, and he also showed that the least squares estimator is the best linear unbiased estimator.

In the late 19th century, Francis Galton used linear regression to study the relationship between parent height and child height. He found that the relationship was approximately linear, and he coined the term "regression" to describe this phenomenon. Galton also developed the correlation coefficient, which is a measure of the strength of the linear relationship between two variables.

In the early 20th century, Karl Pearson further developed the theory of linear regression. He introduced the concept of multiple regression, which allows for the simultaneous modeling of the relationship between a dependent variable and multiple independent variables. Pearson also developed several statistical tests for assessing the significance of the results of a linear regression model.

Linear regression is a widely used statistical method that has applications in a wide variety of fields, including economics, finance, marketing, and the social sciences. It is a powerful tool for understanding the relationship between variables and for making predictions about future values.

Here are some of the key events in the history of linear regression:

1805: Adrien-Marie Legendre publishes the method of least squares.

1809: Carl Friedrich Gauss independently develops the method of least squares.

1877: Francis Galton uses linear regression to study the relationship between parent height and child height.

1900: Karl Pearson introduces the concept of multiple regression.

1903: Pearson develops the Pearson product-moment correlation coefficient.

1936: Ronald Fisher publishes The Design of Experiments, which includes a discussion of linear regression.

1950: John Tukey publishes Exploratory Data Analysis, which introduces the concept of robust regression.

1979: Gary King publishes Unifying Political Methodology, which provides a comprehensive overview of linear regression.

Linear regression is a powerful statistical tool that has been used for over 200 years to understand the relationship between variables and to make predictions about future values. It is a widely used method in a wide variety of fields, and it continues to be an area of active research.

BASIC TERMINOLOGIES

Here we include some basic knowledge and Terminologies that are included or required to understand the project completely.

1. Response variable:
   The variables that are to be estimated are called Response variables. They are also known as Dependent variables, Target variables and Output variables. They are always Continuous in nature. Some examples of Response variables include:
   - House price
   - Taxi fare
   - Price of product
   - Salary of a person
   - Sales in numbers etc.

2. Predictor variable:
   The variables by which the Response variables is/are to be estimated, are called Predictor Variables. They are also known as Independent variable, Features, Covariates etc. They can be Categorical (having different levels) or Continuous in nature. Some examples of Predictor variables include:
   - Number of rooms
   - Distance of a path
   - Gender
   - Education level etc.

3. SLRM:
   Simple Linear Regression Model, also known as SLRM is considered when there is only one Predictor variable that estimates the Response variable. It is the most general form of Linear Regression Model.

4. MLRM:
   Multiple Linear Regression Model, also known as MLRM is considered when there are more than one Predictor variables that estimates the Response variable. It is slightly complex than the Linear Regression Model.

5. Example of SLRM & MLRM:
   Suppose we consider a data set, in which the variable Y is to be estimated (Response variable) with the help of n number of predictor variables given as $x_1, x_2, x_3, \dots, x_n$. Then the MLRM can be given as:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n$$

   Where $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ are the coefficient of estimates.
   Now when we consider just one Predictor variable out of this n number of Predictors then the model that is build is called SLRM. Here β_0 is called the intercept and the other estimates determine the slopes.

6. Correlation Coefficient Matrix:
   This is a square matrix that contains the respective correlation of the Predictor variables. Keep in mind that this matrix is always symmetric with all diagonal elements are equal to 1. The value of this matrix always lies between -1 and +1.

   Now we will discuss what this value means.
   - +1 means there is a strict positive correlation between two variables.
   - -1 means there is a strict negative correlation between two variables.
   - 0 means the two variables under consideration are not correlated at all.

Now we will show the theoretical formula to find this correlation coefficient. Suppose x_1, x_2 be two Predictor variables. Then the correlation coefficient is given by,

$$\rho = \frac{\widetilde{x_1}^T . \widetilde{x_2}}{||\widetilde{x_1}||\,||\widetilde{x_2}||}$$

Where $\widetilde{x_1}$ and $\widetilde{x_2}$ are mean centered Predictor variables.

7. VIF:

   Variance Inflation Factor also called as VIF is a well know process, used to detect the correlation between the Predictor variables. In general, depending on its values we conclude as following:

   - When VIF is small (< 0.1), then we conclude that the Predictor under consideration is highly correlated with other Predictors.
   - When VIF is large (> 10), then we must do additional studies on the Predictor under consideration.

BASIC CONCEPTS

Here we will discuss some basic concepts that help us understand the concept more easily. In this section we will discuss only two topics. They are as follows:

1. Making a Linear Regression Model:

   Now making a model is very simple using R studio. In R we can create a SLRM or MLRM which ever we want by simply using the lm() function. The function lm() stands for Linear Model. The code includes the name of the data set followed by the Response variable name then an "~" sign and then the name of Predictor variable in case of SLRM. However, in case of MLRM the Predictor variables are written together by "+" sign between them.

   Here we discuss the above information with an example. Say, Price is the Response variable, and we consider Density, Oil_Percentage, Hardness, Crispy, Fracture as Predictor variables. Now to make a Simple Linear Regression Model suppose we consider the Density Variable as Predictor. Then the code for this will be:

```
model = lm(data= food_data , Price ~ Density)
summary(model)
```

Now suppose we want to build a MLRM. Then the code for this model will be:

```
model= lm(data=food_data, Price ~ Density+Hardness+Oil_Percentage+Crispy+Facture)
summary(model)
```

2. Linear Regression Models For Categorical Variables:
   The Response variable in Linear Regression Model is always Continuous in nature. But we can consider a Linear Regression Model where we have Categorical Predictor variables. In such case the result includes what is called as Dummy variables. These variables are generated due to the different levels of the categorical variable under consideration. The level that comes first according to alphabetical order is set as reference level and 0 is assigned to it. For the other levels 1 is assigned corresponding to different dummy variables.
   Let us understand this by an example. Suppose we consider a data set that has a categorical variable called Transport. It has 4 levels such as Bus, Auto, Car and Train. [To know the levels of a categorical variable levels() function is used]. If we create a Linear Regression Model, then it will include $(4 - 1) = 3$ Dummy variables. The name of Dummy variables will be TransportBus, TransportCar and TransportTrain. The fourth one TransportAuto will be automatically set as a reference level and a value will be

assigned to it, which is 0. We can observe these different values associated with the variables by using the contrasts() function.

Here we give the table for dummy Encoding.

	TransportBus	TransportCar	TransportTrain
Auto	0	0	0
Bus	1	0	0
Car	0	1	0
Train	0	0	1

INTERPREATATION OF MODELS

1. Model-1:
   In the first model we build a Simple Linear Regression Model of Price as Response variable and Oil_Percentage as Predictor variable. The results give the intercept value 1580.41 and the slope value -14.97. So, the model can be written as,
   $$\widehat{Price} = \widehat{\beta_0} + \widehat{\beta_1} * Oil\_Percentage$$

   Now putting the values of the coefficients, we can write the model as:
   $$\widehat{Price} = 1580.41 - 14.97 * Oil\_Percentage$$

   Interpretation
   For 1-unit increase in Oil_Percentage, the price of the product decreases by Rs.14.97

2. Model-2:
   In the second model we build a Simple Linear Regression Model of Price as Response variable and Density as Predictor variable. The results give the intercept value 733.7651 and the slope value 0.2062. So, the model can be written as,

$$\widehat{Price} = \widehat{\beta_0} + \widehat{\beta_2} * Density$$

Now putting the values of the coefficients, we can write the model as:

$$\widehat{Price} = 733.7651 + 0.2062 * Density$$

Interpretation

For 1-unit increase in Density, the price of the product increases by Rs.0.2062

3. Model-3:

   In the third model we build a Simple Linear Regression Model of Price as Response variable and Crispy as Predictor variable. The results give the intercept value 2144.72 and the slope value -71.33. So, the model can be written as,

   $$\widehat{Price} = \widehat{\beta_0} + \widehat{\beta_3} * Crispy$$

   Now putting the values of the coefficients, we can write the model as:

   $$\widehat{Price} = 2144.72 - 71.33 * Crispy$$

   Interpretation

   For 1-unit increase in Crispy, the price of the product decreases by Rs.71.33

4. Model-4:

   In the fourth model we build a Simple Linear Regression Model of Price as Response variable and Hardness as Predictor variable. The results give the intercept value 2249.1325 and the slope value -7.2255. So, the model can be written as,

   $$\widehat{Price} = \widehat{\beta_0} + \widehat{\beta_4} * Hardness$$

Now putting the values of the coefficients, we can write the model as:

$$\widehat{Price} = 2249.1325 - 7.2255 * Hardness$$

Interpretation

For 1-unit increase in Hardness, the price of the product decreases by Rs. 7.2255

5. Model-5:

   In the fifth model we build a Simple Linear Regression Model of Price as Response variable and Fracture as Predictor variable. The results give the intercept value 928.39 and the slope value 18.91. So, the model can be written as,

$$\widehat{Price} = \widehat{\beta_0} + \widehat{\beta_5} * Fracture$$

   Now putting the values of the coefficients, we can write the model as:

$$\widehat{Price} = 928.39 + 18.91 * Fracture$$

   Interpretation

   For 1-unit increase in Fracture, the price of the product increases by Rs.18.91.

6. Model-6:

   In the sixth model we build a Multiple Linear Regression Model by taking Price as Response variable and the other variables such as Oil_Percentage, Density, Crispy, Hardness and Fracture as Predictor variables. Say, coefficients of the estimates are $\beta_1, \beta_2, \beta_3, \beta_4$ and β_5 respectively. Then the model is given as:

$$\boxed{\begin{aligned}\widehat{Price} = \widehat{\beta_0} + \widehat{\beta_1} * Oil\ Percentage + \widehat{\beta_2} * Density + \widehat{\beta_3} * Crispy \\ + \widehat{\beta_4} * Hardness + \widehat{\beta_5} * Fracture\end{aligned}}$$

After getting the results we observe the values of the coefficients are given as: 1811.3143 (intercept), -15.3751, 0.3539, -7.6961, -9.0464 and -6.9299. Therefore, we can write the model as:

$$\widehat{Price} = 1811.3143 - 15.3751 * Oil\ Percentage + 0.3539 * Density - 7.6961 * Crispy - 9.0464 * Hardness - 6.9299 * Fracture$$

Interpretation

- Keeping Density, Crispy, Hardness and Fracture unchanged, for 1-unit increase in Oil_Percentage, the Price of the product will decrease by Rs. 15.3751
- Keeping Oil_Percentage, Crispy, Hardness and Fracture unchanged, for 1-unit increase in Density, the Price of the product will increase by Rs. 0.3539
- Keeping Density, Oil_Percentage, Hardness and Fracture unchanged, for 1-unit increase in Crispy, the Price of the product will decrease by Rs. 7.6961
- Keeping Density, Crispy, Oil_Percentage and Fracture unchanged, for 1-unit increase in Hardness, the Price of the product will decrease by Rs. 9.0464
- Keeping Density, Crispy, Hardness and Oil_Percentage unchanged, for 1-unit increase in Fracture, the Price of the product will decrease by Rs. 6.9299

Note:

When we construct the Simple Linear Regression Model and compare its values with the Multiple Linear Regression Model, then we see a dramatic change in the values of the predicted estimators. This implies that the Predictor variables are confounders in nature.

7. Model-7:

   Before discussing this model let us discuss about correlation coefficient matrix and Variance Inflation Factor. The matrix gives the correlation between any two pairs of Predictor variables. It values can lies between -1 and +1. We observe that the matrix is symmetric about the diagonal while all diagonal elements are exactly 1. From the matrix we get in the project we observe that the correlation between Oil_Percentage and Crispy is the highest 0.59.

   Again, to detect which is more problematic between Oil_Percentage and Crispy we find out the VIF values for each pf the Predictor variables and find that Crispy has the highest values approximately close to 5.5.

   So, in model number 7 we build a Multiple Linear Regression Model Price as Response variables and Density, Oil_Percentage, Fracture and Hardness as Predictor variables. We observe that the result (the values of the estimators) changed dramatically.

8. Model-8:

   In model number 8 we try to introduce the concept of Collinearity in the Predictor variables to observe what happens in the result. To do this we create a column called Density_in_hundred which is nothing but Density/100. So, the columns Density and the column Density_in_hundred are linearly dependent to each other.

   Now we build a Multiple Linear Regression Model with Price as Response variable and Density, Hardness and Density_in_hundred as Predictor variables. In result we observe that the values corresponding to the Density_in_hundred column is given as NA.

9. Model-9:
   In this Model we will discuss what happens when we consider a Categorical variable in the Predictor variable. To do this we create a new column named Costly and assigned its values as Yes and No with respect to the mean Price of Product. Here Yes and No are the two levels of the variable. As N comes before Y in alphabetical order, so No is set as reference level. The new Dummy variable CostlyYes will come in the result.

   The model can be given as:

$$\widehat{Price} = \widehat{\beta_0} + \widehat{\beta_2} * Density + \widehat{\beta_4} * Hardness + \widehat{\beta_6} * CostlyYes$$

   The predicted values of the estimators are given as 1110.05916, 0.40080, -7.30285, -7.97020. So, the model can be rewritten as:

$$\widehat{Price} = 1110.05916 + 0.40080 * Density - 7.30285 * Hardness - 7.97020 * CostlyYes$$

   Interpretation
   When the product is Not Costly then the value of CostlyYes will become 0 and then we can conclude that for 1-unit increase in Density, the price will be increased by Rs. 0.40080, keeping Hardness unchanged.

SOME POINTS TO REMEMBER

Here we will discuss some concepts that we come across this in detailed discussion section. Here we notice that whenever we write how a model will look like we put the Response variable Price under a “hat”. This is because the Price we will get from computing the model is not the true Price. It is just a predicted Price. On the other hand, the coefficient of estimators denoted by $\beta_i \; 0 \leq i \leq n$ are also predicted values. So, they must be written under the “hat”.

In general, the Predictors variables are always written in capital letters (for example say X_1, X_2 be two predictors). When we decide the sample then they can be written as x_1, x_2.

REFERENCES:

https://en.wikipedia.org/wiki/Linear_regression