Assignment 3

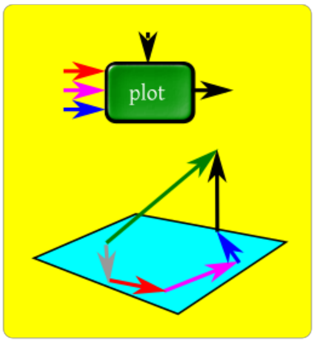
Aytijhya Saha
Roll No. BS2002

Indian Statistical Institute, Kolkata

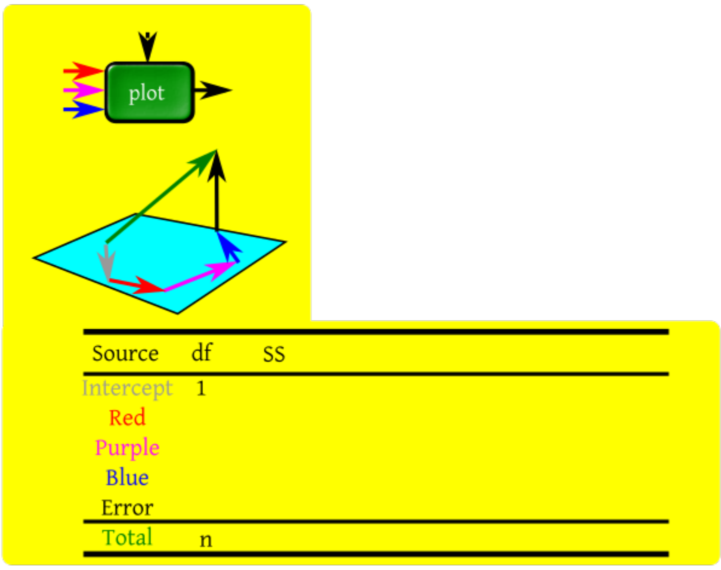
October 9, 2022

ANOVA table

Early statisticians had developed a particular form, using which we can express the splitting up of the sum of squares, i.e the analysis of variance. In order to understand that, we consider the example, where we have three factor inputs and we assume that we have the necessary orthogonality relationship.

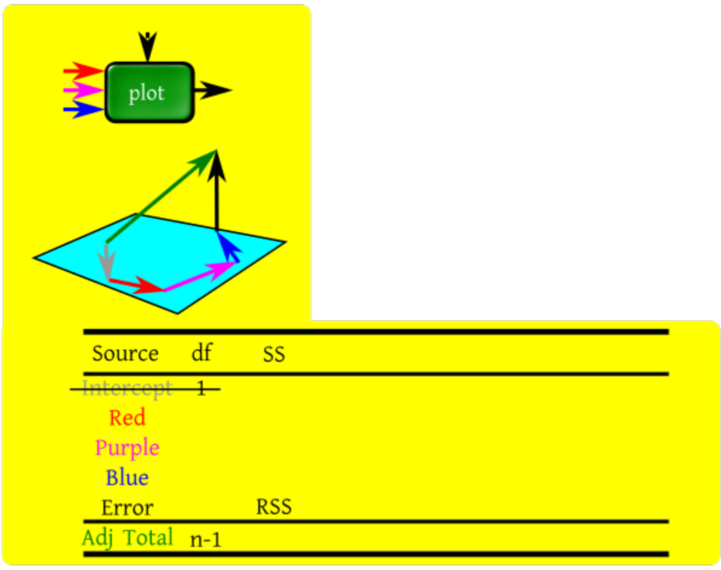


This situation is expressed using a table. We know that the entire space, $\mathbb{R}^n$ can be split up into various mutually orthogonal subspaces- intercept contributes one, each input variable contributes one subspace and the last subspace is devoted to the orthogonal complement of the column space of the design matrix, X. Each row of the table corresponds to one of the subspaces. And the first column is the source column, which consists of the names of different sources of sum of squares- intercept, input 1, ... , input p and the residual. This column will be followed by the degrees of freedom column, where we present the degrees of freedom of the sources, i.e, the dimensions of the corresponding subspaces. Now, we know that total degree of freedom is n and the degree of freedom corresponding to the intercept is 1.



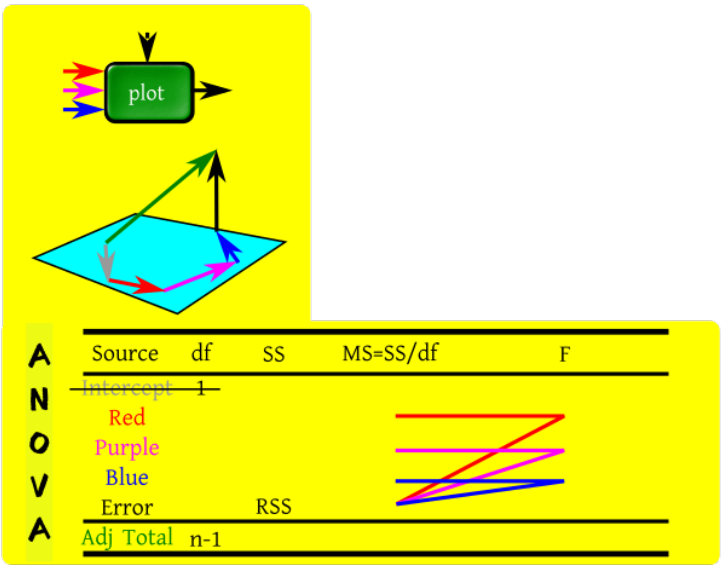
Source	df	SS
Intercept	1	
Red		
Purple		
Blue		
Error		
Total	n	

Traditionally, people do not write the row corresponding to the intercept because the intercept really does not correspond to any of the inputs, whereas all other sources correspond to some arrow in the blackbox. So, we subtract that from the total and the new total will be called the adjusted total, with degree of freedom $n-1$.



Source	df	SS
~~Intercept~~	~~1~~	
Red		
Purple		
Blue		
Error		RSS
Adj Total	n-1	

In the third column of the table, we have the sum of squares due to the sources, which are the squared norms of the different components as shown by the red, pink, blue and black arrows in the figures. In the next column, we have the mean squares, i.e, the sum of squares divided by the corresponding degrees of freedom. There last column consists of the values of the F statistic, which is nothing but the mean square divided by the MSE (mean square of error). Note that this MSE is the unbiased estimator of the σ^2 in the Gauss Markov set-up, we discussed earlier. We use the F statistics to test whether an input has any significant influence on the output or not. If the value of the F statistics is very low, we say that the corresponding input has no significant effect on the output. Obviously, there is no F statistics corresponding to the error because the mean square of error is the acceptable level of error and we are comparing all other sum of squares in terms of that.



Source	df	SS	MS=SS/df	F
~~Intercept~~	~~1~~			
Red				
Purple				
Blue				
Error		RSS		
Adj Total	n-1			

This table is known as ANOVA table, that was motivated completely from the viewpoint of the set-up where we have only factor inputs. For a long time, the ANOVA table was considered to be the most important output of the linear model analysis.