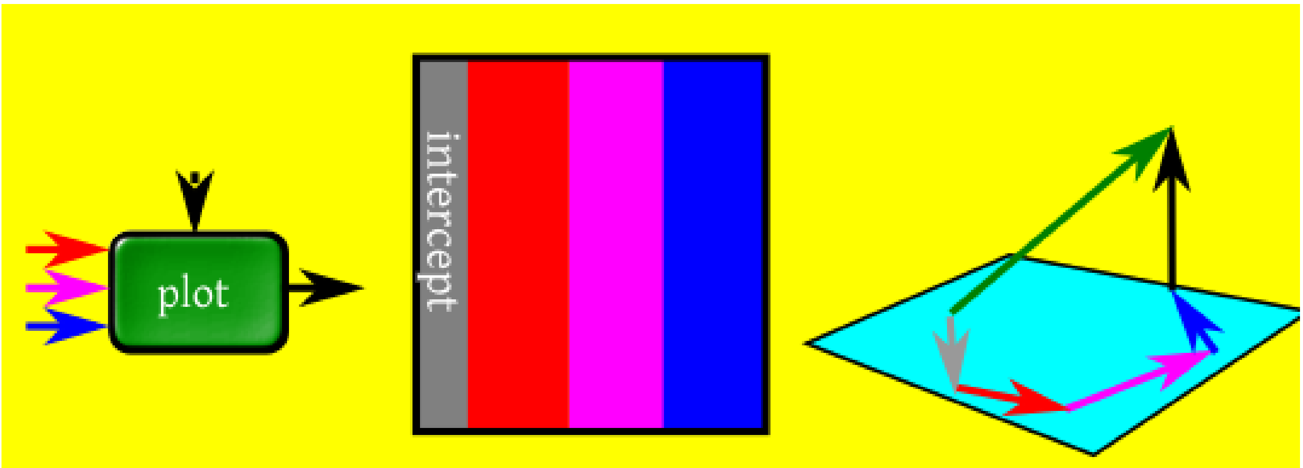
Anova Table: Geometry

Arunav Bhowmick (BS2025)

October 2022

We saw that the condition under which we could split the total variation as seen in the output into component wise variation ascribable to the different inputs was 'loose' orthogonality of the columns. This splitting may be explained in terms of our familiar projection idea.



In the diagram, we have the column space of X and the data vector $\vec{y}$, represented by the green arrow. What we are doing here, is that, we are splitting up the data vector into lots of orthogonal components. So, we are expressing $\vec{y}$ as the sum of five mutually orthogonal vectors, as represented by the gray, red, purple, blue and black arrows. Since we can only show this diagram in three dimensional space where we cannot come up with five mutually orthogonal directions, take this diagram with a pinch of salt.

Each of these vectors is actually the projection of $\vec{y}$ along the respective direction. To elaborate, we project the green arrow along the direction of the intercept, i.e., along the direction of the vector (1,1, ...,1) to obtain the gray arrow. After we have removed this part, we project the remainder along the direction outlined by the red columns after removing the effect of the intercept, to obtain the red arrow. Hence, the gray and red arrows are perpendicular. Similarly, we do this for the purple and blue things to get the purple and blue arrows respectively. These completely explain the projection of $\vec{y}$ onto column space of X. Whatever remains is orthogonal to these vectors, and that is the black arrow.

So in this case, by a generalised version of Pythagoras theorem, we can say that the norm square of the data vector is the sum of the norm squares of these orthogonal components. These norm squares of each of the components play the role of the component sum of squares and they also add up. So early statisticians figured out that this is the necessary and sufficient condition for which we can split up the total sum of squares in terms of the component sum of squares. Unfortunately, they also found that this is not always possible.