Hypothesis Testing : Test statistic

Prisha Reddy Bobbili

September 2022

"The column space of X_1 is a sub space of the column space of X.However the columns of X_1 need not be a subset of the columns of X."

Let us understand this through the following example :
Consider the scenario that we have been using in the previous examples. Three varieties of seeds are compared , and the linear model also consists of an intercept term μ. Call the design matrix in this case is X.

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

Our main goal is to see the differences in the yield obtained by variety 1 and variety 2. Consider the null hypothesis, H_0, as follows:

$$H_0 : (\alpha)_1 = (\alpha)_2 \tag{1}$$

$$H_1 : (\alpha)_1 \neq (\alpha)_2 \tag{2}$$

In simple words, we test to see if the yields are really different, with our null hypothesis being, that they are not.

Thus, a new design matrix X_1 is constructed with the given information as for the previous design matrix, but also including the new additional information.We shall now have a single column associated with varieties 1 and 2.

$$\alpha_1 = \alpha_2$$
$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

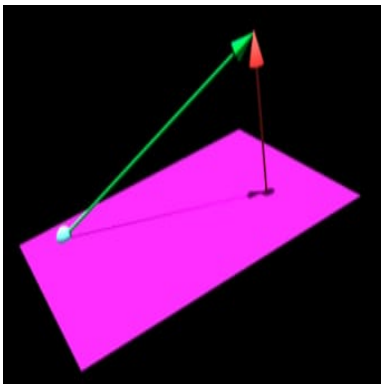
Thus, X_1 can be visualised as follows:
The first and last columns from the previous design matrix are unchanged.The remaining two columns are combined (more specifically, added) to form a single column.

Compare the two design matrices.Clearly the column space of X_1 is a subspace of the column space of X. However, the second column of X_1 is not a column of X.

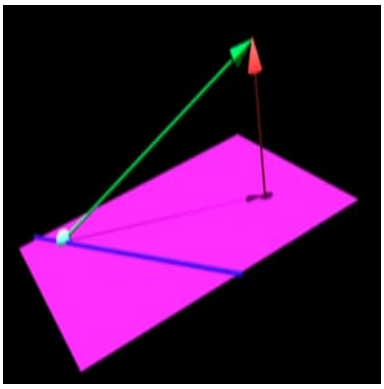
The column space of X_1 is a sub space of the column space of X.However the columns of X_1 need not be a subset of the columns of X.

1 Understanding The Pictures

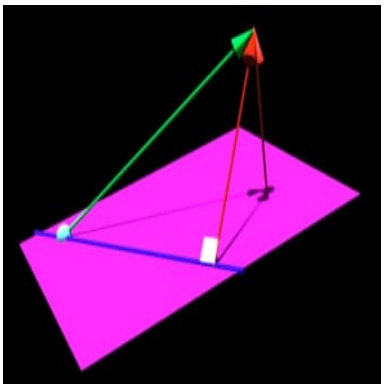
We shall now develop the basic ideology of how the test of hypothesis is conducted in such cases. The mathematical part of the tests will follow later.
Consider a design matrix X.Visualise the column space of X to be the pink plane.
The green vector represents the data provided or given.
The orange vector, thus, is the residual vector.



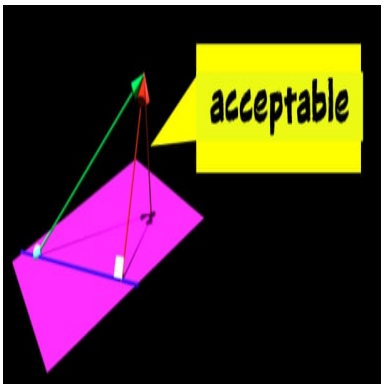
The subspace of this column space of X (i.e.the column space of X_1) is represented by a line passing through the origin.In the given picture it is coloured blue.



Our next step would be to fit this restricted model. Thus, we drop a perpendicular from the head of the green vector (data) onto the blue line (column space of X_1), to obtain a new residual vector. This vector represents the error obtained by adopting the restricted model.



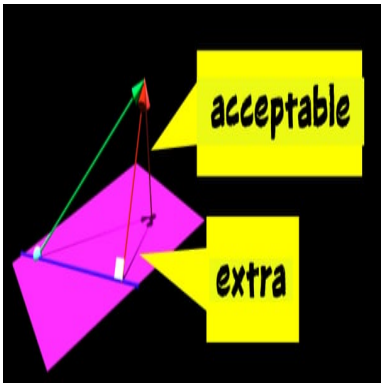
Thus, the original residual vector, which represents the error obtained naturally, or without the additional restriction, is considered acceptable.



What we test is if the new residual vector obtained represents an error close enough to that obtained by the "acceptable" residual vector or not. For this, we consider the extra amount of error obtained by the new residual vector.

The "extra" error vector is shown as the shadow of the new residual vector cast over the original plane (since we are considering the squared norm).

Note that, since the error is obtained with additional restrictions, the amount of error is bound to increase (or at least stay unchanged). This can also be explained in the following figure, where a nice little right triangle is formed by the "acceptable" residual vector, the new residual vector, and the "extra" error vector, with the new residual vector being the hypotenuse. Thus, pythagoras theorem helps us visualise this even in higher dimension.



Consider the following scenario: A certain task demands an additional Rs.50 for the same amount of work as before. Suppose person A earns only 100 rupees per month, while person B earns 1,00,000 rupees per month. Obviously, it is so much more convenient for person B to manage with the additional charge that it is to person A.

You might be ready to pay an additional charge of Rs.500 on an air ticket that was originally priced at say around 30,000 rupees, but might not want to spend even 5 rupees more on much cheaper things, for say bananas. Therefore, we can say, not just the extra error, but also the acceptable amount of error is crucial in the testing in such cases.