Hypotheses Testing: General Subspace Idea

Arunav Bhowmick (BS2025)

September 2022

We recall the example that we were dealing with previously. It involved the two classes of models described below.

$$\begin{gathered} y_{ij} = \mu + \alpha_i + \epsilon_{ij} \\ i = 1, 2, 3, \quad j = 1, 2 \\ \epsilon_{ij} \sim (0, \sigma^2), \quad \sigma^2 > 0 \\ \mu, \alpha_1, \alpha_2, \alpha_3 \in \mathbb{R} \end{gathered} \tag{1}$$

and,

$$\begin{gathered} y_{ij} = \mu + \epsilon_{ij} \\ i = 1, 2, 3, \quad j = 1, 2 \\ \epsilon_{ij} \sim (0, \sigma^2), \quad \sigma^2 > 0 \\ \mu \in \mathbb{R} \end{gathered} \tag{2}$$

There is something special about these classes of models in the sense that (2) is not only a subset of (1), but is in fact its subspace. To understand this, we consider their design matrices. (1) has the design matrix

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \tag{3}$$

whereas (2) has the design matrix

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \tag{4}$$

It can be easily verified that column space of (4) is a subspace of column space of (3).

In general, we will be given two models. These are described below.

$$\begin{gathered} \vec{y} = X\vec{\beta} + \vec{\epsilon} \\ \vec{\epsilon} \sim N(\vec{0}, \sigma^2 I), \quad \sigma^2 > 0 \\ \vec{\beta} \in \mathbb{R}^p \end{gathered}$$

and,

$$\begin{gathered} \vec{y} = X_1\vec{\beta} + \vec{\epsilon} \\ \vec{\epsilon} \sim N(\vec{0}, \sigma^2 I), \quad \sigma^2 > 0 \\ \vec{\beta} \in \mathbb{R}^q \\ \mathcal{C}(X_1) \subseteq \mathcal{C}(X) \end{gathered}$$

Note that we have made one more assumption here. We assume that the errors are not only mean $\vec{0}$ and variance $\sigma^2 I$, but are also jointly normal. This is a typical assumption under which we perform our tests of hypotheses because we need the distribution of the test statistic. So, we need a full distributional assumption and the normal distribution assumption is the most natural one since it makes the mathematics simple.

The first set of models, which basically gives us a family of distributions, is believed to be a good fit. What it means is that the true distribution of our data lies in this family of distributions. So, $\vec{y}$ will lie pretty close to column space of X. Thus, we assume based on whatever goodness of fit we have done, domain knowledge, diagnostics, etc., that we have a good fit.

Now, someone gives us a restricted set of models which is same as the first one except that we now have X_1 in place of X. X_1 need not be made up only of columns chosen from that of X. It may be an entirely different design matrix. As we have mentioned earlier, $\vec{\beta}$ is just an indexing parameter. So, $\vec{\beta}$ in the first and second models is possibly different. In the first model, $\vec{\beta}$ lies in $\mathbb{R}^p$ and in the second model, it lies in $\mathbb{R}^q$. That is, we assume that X is order $n \times p$ and X_1 is order $n \times q$. n remains the same as our observation vector is the same. We also assume that column space of X_1 is inside column space of X. That is what we mean when we say that the second model is a subspace of the first model.

One simple way to visualise this is: it is possible to put restrictions on $\vec{\beta}$ of the first model such that the first model is reduced to the second model. As in our last example, if we put the restriction that all the α_i's are equal, then (1) gets reduced to (2) and that is a simple proof that the second model is indeed a subspace of the first model.