ANOVA TABLE : INTERPRETATION

Rajarshi Biswas

Roll-BS2011

INDIAN STATISTICAL INSTITUTE

Suppose , we are estimating for yields of a crop from some (say m) varieties of fertilizers .
Let us define :

y_{ij} as yield produced by j-th plot of the i-th variety .

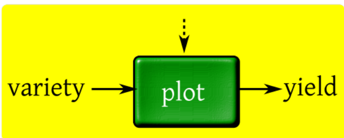
$WSS = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y_{i.}})^2$

$BSS = \sum_{i=1}^{m} n_i (y_{i.} - \bar{y})^2$

$TSS = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y_{..}})^2$

where , $\bar{y_{..}} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{m} n_i}$ and $\bar{y_{i.}} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}, \forall i \in \{1, 2, 3, ..., m\}$

Let us understand this with an example .



Here, TSS(total sum of squares) is overall effect observed on the yield , BSS(Between sum of squares) is the effect observed due to different variety ,WSS(Within sum of squares) is the effect observed due to random error .

Now , we will see how to prove $TSS = BSS + WSS$,i.e, the total variation in the output comes from the input and there is no extra variation within the model .

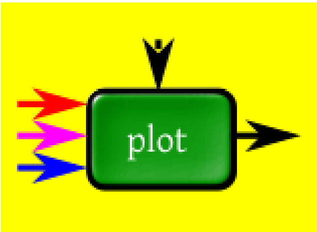
$TSS = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y_{..}})^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y_{i.}} + \bar{y_{i.}} - \bar{y_{..}})^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y_{i.}})^2 + 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{y_{i.}} - \bar{y})(y_{ij} - \bar{y_{i.}}) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{y_{i.}} - \bar{y})^2$

Now ,$2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{y_{i.}} - \bar{y})(y_{ij} - \bar{y_{i.}}) = 2 \sum_{i=1}^{m} (\bar{y_{i.}} - \bar{y}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y_{i.}}) = 2 \sum_{i=1}^{m} (\bar{y_{i.}} - \bar{y})(\sum_{j=1}^{n_i} y_{ij} - \sum_{j=1}^{n_i} \bar{y_{i.}}) = 2 \sum_{i=1}^{m} (\bar{y_{i.}} - \bar{y})(n_i \bar{y_{i.}} - n_i \bar{y_{i.}}) = 0$

So , $TSS = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y_{i.}})^2 + 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{y_{i.}} - \bar{y})(y_{ij} - \bar{y_{i.}}) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{y_{i.}} - \bar{y})^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y_{i.}})^2 + 0 + \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{y_{i.}} - \bar{y})^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y_{i.}})^2 + \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{y_{i.}} - \bar{y})^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y_{i.}})^2 + 0 + \sum_{i=1}^{m} n_i (\bar{y_{i.}} - \bar{y})^2 = WSS + BSS$

So , we proved $TSS = WSS + BSS$ total variation in the yield is variation due to different varieties and variation due to random error . This interpretation of splitting the variation into 2 different components and analysing the variation of the yield is what we study under the ANOVA(Analysis of Variance).

Here , in this model we have split the TSS as a sum of WSS and BSS . Now , what if we have more than one variable ?



Here we have more than 1 input . So , we will have more terms of BSS . Like in this model we have

$TSS = BSS_1 + BSS_2 + BSS_3 + WSS$

where BSS_i is the variablity due to the i-th variable .
