Computing the variance component estimates and effective degrees of freedom

Why am I doing this?

The main goal of my PhD project is to develop a general framework for designing two-phase MudPIT-iTRAQ experiments. The purpose of this framework is to find the optimised designs specifically for MudPIT-iTRAQ experiments. A paper produced by Jarrett and Ruggiero (2008) demonstrated that despite having the same number of biological and technical replicates, the outcome of an experiment can be affected by the type of design used. While this paper used microarrays, MudPIT-iTRAQ experiments have their own set of unique problems. Moreover, in order to develop this general framework, the methods for data analysis are equally important, because different methods may generate different results and affect our perception of an optimised design. Hence, different methods for analysing MudPIT-iTRAQ experiments also need to be investigated.

To determine which method is better, the estimates of the variance components can be used to for comparison between the two methods.

What I am going to talk about – summary

I will discuss the theory of …..

Compare to Jarrett and Ruggiero (2008) paper – satterthewaite approximation

Discuss the theory I have here which is the generalisation of the theory described in their paper.

Present the R function getVcEDF() that implements the theory

simData()

EDFplot()

Present the theory

Present function

Jarrett and Ruggiero (2008)

What is variance components (VC) estimates and effective degrees of freedom (EDF)

Variance components (VC) estimates can be generated from the linear combination of the expected mean squares, then use the calculated mean squares based on the experimental data. Another method of estimating variance components is using Fisher’s scoring algorithm

EDF is used to assess the effectiveness of the variance component estimates.

To compute the VC and EDF we need to define the Score function and Fisher’s information matrix.

* Why – VC can be estimated by the Fisher’s information matrix and score function using the scoring algorithm
* EDF can be computed by Satterthwaite approximation which is the twice of square of the mean divided by the variance. The variance can be obtained from the Fisher’s information matrix.

Theoretical ANOVA table can be derived based on the design of the experiment. From the theoretical ANOVA table the EMS can be obtained. The EMS and the calculated mean squares (MS) based on the given data can then be used to estimate the variance components estimates.

The model of the analysis can also be derived from the design of the experiment. Once the data set is generated from the experiment, we can either perform the ANOVA method, to obtain the mean squares, or REML method, to estimate the variance component by Fisher’s scoring. However, in Richard and Kathy’s paper, the REML estimates used the mean squares and degrees of freedom from the ANOVA table, hence, the ANOVA method is still required to perform initially.

To get the score function and fisher’s information matrix, we need to define the likelihood function.

Mean squares of the ANOVA has Chi-square distribution, hence the likelihood function is….

The score function which is also the first derivative with respect to expected mean squares is ……

The expected Fisher’s information matrix is the expected value of then the second derivative of the likelihood function which is DF divided by the twice of the square of the mean squares.

Hence, expected Fisher’s information matrix can be constructed by the degrees of freedom and mean squares.

With the expected Fisher’s information matrix and score function, we can perform the scoring algorithm to obtain the better estimates of the mean squares.

However, what we interested is the variance components that make up the expected mean squares.

A matrix which is constructed based on the variance components’ coefficients of the expected mean square can be used to transform the expected Fisher’s information matrix and score function from with respect to the mean squares to with respect to the each individual variance component. This is also known as change of variable technique.

The variance components can then be optimised using the scoring algorithm. The initial values can be any value.

Once the variance components are optimised, the EDF can be computed using the Satterthwaite approximation. That is twice of square of the mean divided by the variances.

The mean obtained from the expected mean squares.

The variance is calculated from the expected fisher’s information matrix.

Since the transformation was applied to the score functions and the expected Fisher’s information matrix from the mean squares to the individual variance components, we need to transform back to the mean squares again to compute its EDF.

Expected mean square is the variance components’ coefficients multiply the optimised variance component estimates.

Variance is obtained from the sum of the elements of the interest of the expected fisher’s information matrix. Note the variance components’ coefficients will also need to be using during the summation, the variance of a sum of two random variables

Where a and b are constant of the variance components’ coefficients. If X and Y are independent random variables, then Cov(X,Y) equals zero.

Only interested in the EDF of the stratum where treatment effect is been estimated.

Function

getVcEDF function does everything above

Example with input and output

EDF plots