Analysing an incomplete paired comparison experiment using multiple regression.

The paired comparison experiment can be regarded as equivalent to a multiple regression. Let there be *n* judgements on *n* pairs, sampled from *k* stimuli. The various vectors and matrices are:

j: *(n x 1)* vector, containing the *n* judgements made by a subject, scaled so that +1 indicates a strong preference for the stimulus on the right and -1 a strong preference for the stimulus on the left.

p: *(k x 1)* vector, containing the preference function, where p_j is the preference for the *j*th stimulus.

D: *(n x k)* matrix, containing the 'design matrix', which says which stimuli are presented in each pair, +1 indicating the stimulus on the right and -1 indicating the stimulus on the right. It should be noted that the rows of this matrix are entirely zero except for the two items contained within the pair.

e: (n x 1) vector of error terms, representing random variation in each judgement. This is presumed in the first instance to be *N(0,σ)* (i.e. normal distribution, mean zero, standard deviation of sigma for all judgements), particularly if six point preference judgements are made. For binary judgements, a binomial error function and logistic regression could be used.

An illustrative example of the matrices for five pairs based on four stimuli, would look thus:

$$\begin{bmatrix} j_1 \\ j_2 \\ j_3 \\ j_4 \\ j_5 \\ j_6 \end{bmatrix} = \begin{bmatrix} 0 & +1 & -1 & 0 \\ +1 & 0 & -1 & 0 \\ 0 & 0 & +1 & -1 \\ +1 & 0 & 0 & -1 \\ 0 & +1 & -1 & 0 \\ -1 & +1 & 0 & 0 \end{bmatrix} \bullet \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}$$

Calculating through, one can see that, ignoring error terms, and for the design matrix given,

$$j_1 = p_2 - p_3$$
$$j_2 = p_1 - p_3$$
$$j_3 = p_3 - p_4$$
etc..

Overall the equation is **j** = **D**.**p** + **e**, which is formally identical to the conventional multiple regression equation, **y**= **X**.**b** + **e**. That means that estimating the values of the preferences, **p**, and their standard errors and other associated statistics, is equivalent to

carrying out a conventional multiple regression (although see below about the origin), and the standard equations and programs can be used. Note in particular that the design does not need to be complete or balanced. All judgements for any possible pairs are used to get maximum likelihood estimates of the **p** vector and its error terms.

Two practical points for implementing the method are that:

i) The design matrix does not have a column marked 'constant', and therefore when using regression programs such as *SPSS* one needs to fit a model 'through the origin', which forces the constant to be zero.

ii) If the design matrix, **D**, is *n x k* then in the multiple regression it is necessary to calculate $\mathbf{D}^T.\mathbf{D}$, of size *n* x *n*, and $\mathbf{D}^T.\mathbf{D}$ is inevitably singular. One should therefore use a reduced **D** matrix, call it **D*,** of size *n* x (*k*-1), where a single column has been omitted. When the model is fitted through the origin, then the preference for that omitted stimulus is fixed at zero, and all preferences for other stimuli are relative to it. It is arbitrary which column is omitted.

iii) The absolute preference levels depend on the choice of column omitted in creating **D***. It therefore makes sense to re-standardise the preference values so that they have a mean of zero.