Creating an X matrix for a Two-way ANOVA

Consider an experiment to determine the effect of diet and exercise on cholesterol levels. Exercise is a categorical variable with two levels, "Yes" and "No"; diet is a categorical variable with three levels: a normal diet ("Normal"), a low fat diet ("Low Fat"), and a diet supplemented with oat bran ("Supplement"). Each categorical variable will be represented in the **X** matrix with one less column than there are levels of that variable; *i.e.* the number of degrees of freedom for each categorical variable is one less than the number of levels for that variable.

Obs.	Exercise	Diet	μ	e_1	d_1	d_2
1	No	Normal	1	1	0	1
2	No	Normal	1	1	0	1
3	Yes	Normal	1	-1	0	1
4	Yes	Normal	1	-1	0	1
5	Yes	Normal	1	-1	0	1
6	No	Low Fat	1	1	1	0
7	No	Low Fat	1	1	1	0
8	Yes	Low Fat	1	-1	1	0
9	Yes	Low Fat	1	-1	1	0
10	Yes	Low Fat	1	-1	1	0
11	No	Supplement	1	1	-1	-1
12	Yes	Supplement	1	-1	-1	-1
13	Yes	Supplement	1	-1	-1	-1

In this example, the following coding was used:

1. For a categorical variable with k levels, $k - 1$ columns (dummy variables) will be entered into the **X** matrix.

2. Consider a single level of a categorical variable, say, i, where $i = 1, \ldots k$. For each level except the kth, set the ith dummy variable equal to 1 if the value for the observation is the ith level of that variable, and set it to 0 otherwise.

3. For the kth level, set the value of all the dummy variables ($i = 1, \ldots k - 1$) to -1.

Note that this particular form of coding is arbitrary; a variety of other methods for creating the dummy variables will produce acceptable matrices. The only requirements are that the comparisons between the different levels which the dummy variables define are equivalent to a test of equality among the means of all of the levels, and that the resulting matrix is of full rank. One advantage of the method presented here is that the individual coefficients which are estimated are easily interpretable; *i.e.* they represent the difference between the effect of each level of the categorical variable and the last level.

The columns of the **X** matrix described above represent the main effects of the two factors of interest. They will allow us to estimate coefficients to test hypotheses about one of the factors, ignoring the level of the other factor. What is often of much greater interest in linear models with categorical variables are the interactions, which are concerned with the differences in the effect of one of the categorical variables across the levels of the other categorical variables. In the current example, tests of the main effects would answer questions like "Ignoring diet, does exercise affect cholesterol levels?", while an interaction could answer questions like "Is the effect of exercise different depending on what kind of diet is used?". Fortunately, the interaction columns can be easily generated from the main effects columns through the use of a Kronecker product.

Definition Given two matrices, **A** and **B** with dimensions $n_1 n_2$, and $n_3 n_4$, respectively, the Kronecker product of **A** and **B**, denoted by $\mathbf{A} \otimes \mathbf{B}$ is an $(n_1 \cdot n_3)(n_2 \cdot n_4)$ matrix composed of $n_1 \cdot n_2$ submatrices, each formed by multiplying **B** by one element of **A**.

For example if,

$$A = \begin{bmatrix} 1 & 4 & 3 \\ 2 & 4 & 5 \end{bmatrix}_{2\times 3} \qquad \text{and} \qquad B = \begin{bmatrix} 3 & 2 \\ 1 & 4 \end{bmatrix}_{2\times 2}$$

then

$$A \otimes B = \begin{bmatrix} 3 & 2 & 12 & 8 & 9 & 6 \\ 1 & 4 & 4 & 16 & 3 & 12 \\ 6 & 4 & 12 & 8 & 15 & 10 \\ 2 & 8 & 4 & 16 & 5 & 20 \end{bmatrix}_{4\times 6}$$

To create interaction columns, we simply use, for each row, the Kronecker product of the two (single row) matrices formed from the columns corresponding to each of the effects from which the interaction is being built. In the diet/exercise example, we could form the interaction columns by taking the Kronecker product of the single column labeled e_1 with the two columns labeled d_1 and d_2, resulting in the following $\mathbf{X}$ matrix:

Obs.	Exercise	Diet	μ	e_1	d_1	d_2	ed_1	ed_2
1	No	Normal	1	1	0	1	0	1
2	No	Normal	1	1	0	1	0	1
3	Yes	Normal	1	-1	0	1	0	-1
4	Yes	Normal	1	-1	0	1	0	-1
5	Yes	Normal	1	-1	0	1	0	-1
6	No	Low Fat	1	1	1	0	1	0
7	No	Low Fat	1	1	1	0	1	0
8	Yes	Low Fat	1	-1	1	0	-1	0
9	Yes	Low Fat	1	-1	1	0	-1	0
10	Yes	Low Fat	1	-1	1	0	-1	0
11	No	Supplement	1	1	-1	-1	-1	-1
12	Yes	Supplement	1	-1	-1	-1	1	1
13	Yes	Supplement	1	-1	-1	-1	1	1