UPPSALA
UNIVERSITET

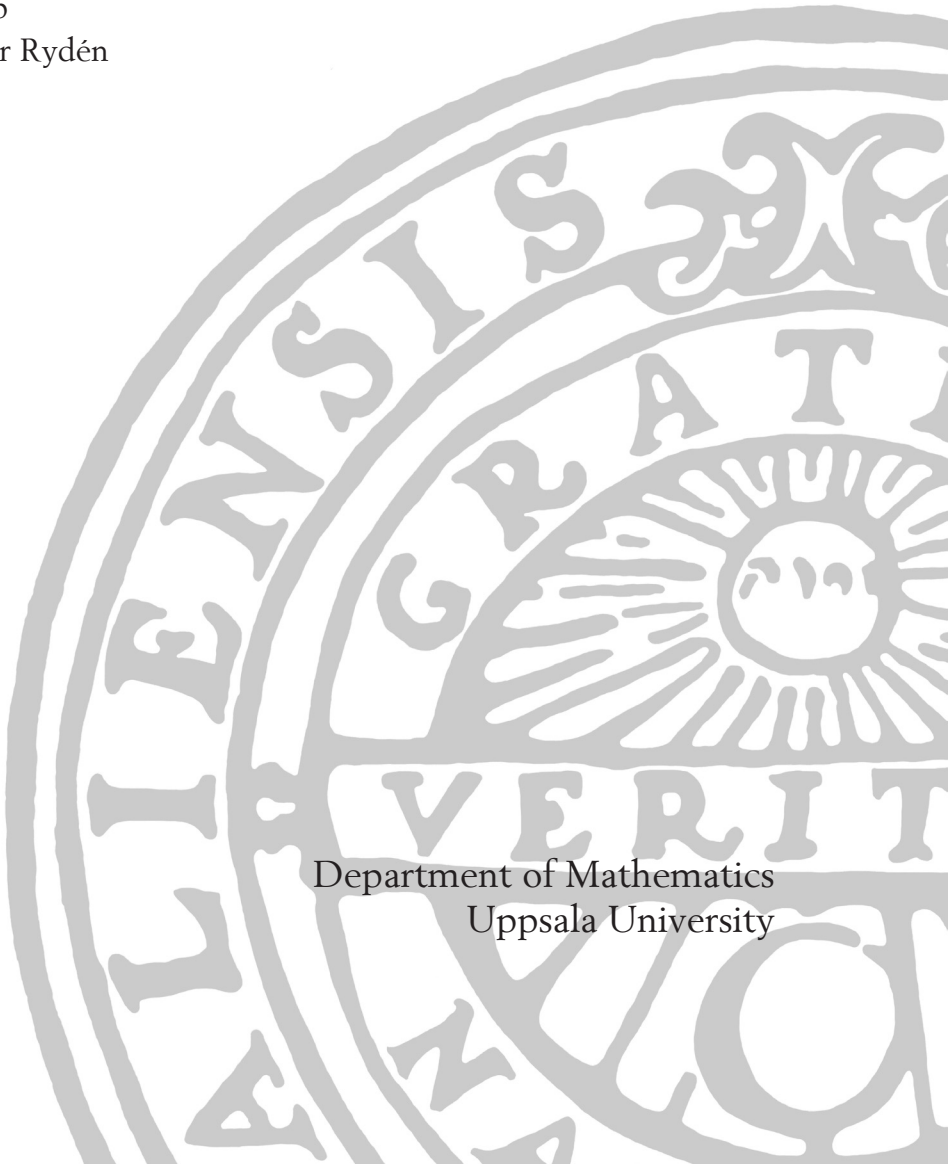
U.U.D.M. Project Report 2010:14

Coding in Multiple Regression Analysis: A Review of Popular Coding Techniques

Stina Sundström

Examensarbete i matematik, 15 hp
Handledare och examinator: Jesper Rydén

Juni 2010



Department of Mathematics
Uppsala University

Coding in Multiple Regression Analysis: A Review of Popular Coding Techniques

Stina Sundström

2010-06-01

Abstract

In previous courses when coded matrices have been mentioned there has been little information on how or why these matrices are coded in the way that they are. The information in the textbooks has not been so detailed. My aim with this rapport is to make a review of this topic. I will discuss how, why and when specific coding methods are used and also mention how these methods can be used in R.

Contents

1 Introduction

In this paper will the choice of different coding schemes be discussed. But what are coding schemes and why are they used? A coding scheme consists of coded variables. Coding variables is a way to change qualitative data into quantitative data. To code a variable is to assign a numerical value to a qualitative trait. Gender is for example a qualitative trait and we can assign female and male with numerical values, say 0 and 1. These may also be refered to as indicator variables. We now have coded variables to work with. A problem when an experimenter makes a design which contains a large amount of data and wants to test some hypothesis, is that it can be difficult to perform calculations when the design matrix is big and complex. It might also be very inefficient since it can be time consuming. To solve the problem the experimenter can use coded variables. This will be much easier to compute and to test hypotheses. One might wonder if there are any limitations of what values to select, can we choose any values that we want? We can, since any pair of different values will yield the same statistical result because correlations are scale free, but it is wise to select values that provide direct estimation of the parameters of interest.[1]

Multiple regression analysis (MRA) is a fundamental technique that can be used for various models, for example ANOVA can be used with MRA. The definition of a linear regression model for a one factor design with k treatment levels is

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

This is a well known equation where only $(k+1)$ independent variables (X_i) are defined in the model to represent all k levels of the treatment.

Let us consider an example where Y is the amount of carbonic acid that remains in a soda, X_1 is time, and the second independent variable is what type of packaging (glass bottle or aluminum can) that is used, we use the coding:

$$X_2 = \begin{cases} 1 & \text{if glass bottle} \\ -1 & \text{if aluminum can} \end{cases}$$

In this case, the first-order linear regression model would be:

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon$$

and has the response function

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

This function is as follows for the two types of packaging

$$E(Y) = (\beta_0 + \beta_2) + \beta_1 X_1$$

$$E(Y) = (\beta_0 - \beta_2) + \beta_1 X_1$$

The first equation represents the packaging glass bottles and the second equation represents the packaging aluminum can. A test whether the regression lines are the same for both types of packaging involves $H_0 : \beta_2 = 0$, $H_1 : \beta_2 \neq 0$. An alternative coding scheme can be made by using 0, 1 as indicator variables. This would give us

$$Y = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon$$

where X_{i1} = time and the indicator variables are

$$X_{i2} = \begin{cases} 1 & \text{if glass bottle} \\ 0 & \text{if otherwise} \end{cases}$$

$$X_{i3} = \begin{cases} 1 & \text{if aluminum can} \\ 0 & \text{if otherwise} \end{cases}$$

The two response functions would then be

$$E(Y) = \beta_2 + \beta_1 X_1$$

$$E(Y) = \beta_3 + \beta_1 X_1$$

To test whether or not these two regression lines are the same would involve the hypotheses $H_0 : \beta_2 = \beta_3$, $H_1 : \beta_2 \neq \beta_3$. Note that the hypothesis H_0 can be formulated as a difference $H_0 : \beta_2 - \beta_3 = 0$. Such linear combinations of parameters is discussed in the following section.

We now turn to coding schemes. A coding scheme is equivalent to the choice of planned ANOVA contrasts. Contrasts are linear combinations of parameters of the form

$$\Gamma = \sum_{i=1}^{a} c_i \mu_i$$

where the contrast constants sum to zero, $\sum_{i=1}^{a} c_i = 0$. For example, say that we have the hypotheses

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

Then the contrast constants are $c_1 = 1$ and $c_2 = -1$.

Let us take a simple example where A, B, C and D are four factors:

A	B	C	D
-1	1	0	0
0	0	-1	1
-1	-1	1	1
-2	1	1	0
-3	1	1	1

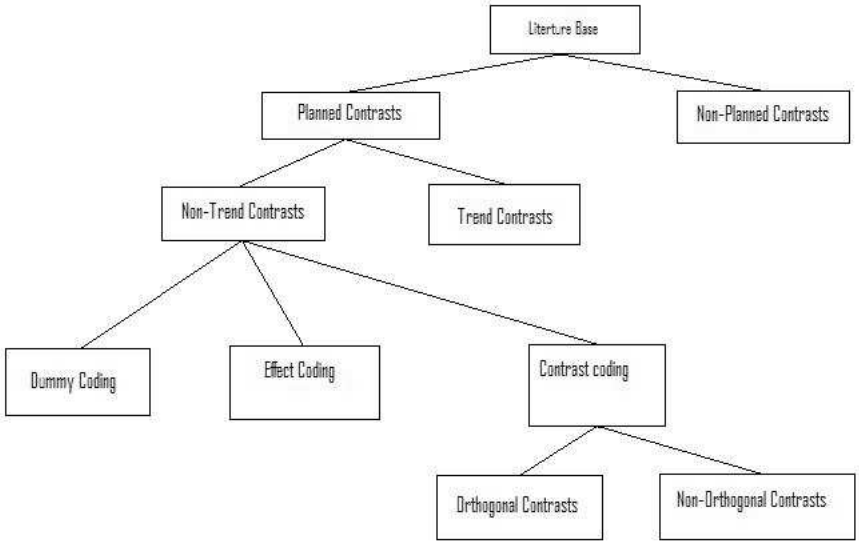


Figure 1: Decision tree for determining appropriate coding structure

In this paper, I will have my starting point of view based on the decision tree for determining appropriate coding structure, which has been published by Davis [2] (see Figure 1). When constructing a coding scheme one must ask oneself what type of hypothesis is to be answered. If previous literature or knowledge allows us to make a specific hypothesis we should examine planned contrasts, if not, we choose non-planned contrasts. The next step is to decide whether to compare group means or study trends of the means. If the trends are desired we create trend or polynomial contrasts to see how the means react across levels. If differences of means are of interest, we must decide what type of mean difference we want to look at. Dummy coding is to prefer if we want to compare a single group against all groups, effect coding is the best choice when we want to test individual groups against each other and if we want to combine both these types of coding, then we use the most complex form of contrast coding. The last step is to determine whether the desired contrasts should be orthogonal or not which will make the result easier to understand.

One coding technique that is worth mentioning but will not be further investigated in this paper is nonsense coding. It is a coding technique where the coded variables are chosen arbitrarily or at random.

2 Planned vs. Unplanned Contrasts

Planned contrasts seem to have a higher status in the literature, unplanned contrasts, also called post hoc contrasts, have been referred to as "data snooping" and "milking a set of results"[3], but unplanned contrasts can also have their benefits. The problem with planned contrasts is that they should not be used

in new areas where the theory yet has not been fully developed, so unplanned are a better choice when you need to coax for the data.

But in most cases, planned contrasts are more useful than unplanned contrasts and have some advantages over unplanned contrasts. Davis in [2] gives the following views.

- Planned contrasts gives increased strength against Type II errors since you do not need to make any irrelevant or unwanted group differences.
- Unplanned contrasts take many or all possible contrasts to consideration which leads to that it must be corrected for all contrasts, even the contrasts that are of no interest.
- When constructing planned contrast you need to be more specific and think through the hypotheses that you want to create.
- The interpretation is easier made since you do not need to interpretate results that are irrelevant or unrelated.

If you choose to use planned contrasts, the next step is to choose trend or non-trend contrasts.

3 Trend vs. Non-Trend Contrasts

A trend contrast does not test if two means are equal but if the mean over the different levels forms a specific pattern. Trend analysis is particularly useful when we want to investigate effects of treatment or other type of data which is not linear and can be discovered with graphs. Such trend contrasts are only possible, according to [2], if

- they are quantitative (e.g., meters of length, minutes of time, liters of water, and so on)
- the levels have the same distance (e.g., 1, 2, 3m)

The model for trend analysis is (according to [4])

$$y_{ij} = \mu + \beta_1(V_1 - \bar{V}) + \beta_2(V_2 - \bar{V})^2 + \beta_3(V_3 - \bar{V})^3 + \ldots + \varepsilon_{ij}$$

where y_{ij} is a polynomial, β_k is the slope, V_i the numerical value of the ith factor level and $\bar{V}$ is the mean of V_i. This equation tests if for each slope in the equation the null hypothesis $\beta_k = 0$. If β_k differs a lot from zero then the null hypothesis is rejected and other nonlinear models will be evaluated. This model is best suited if there are good reasons to believe that the means are related to the levels by a linear function according to [4]. Trend contrasts are also referred to as polynomial contrasts by [2]. This type of contrast requires correct weighting of all treatment means to get correct trend estimation.

Three questions are to be considered in the trend analysis concerning the function relating the independent and the dependent variable:

1. if the data shows an overall trend upwards or downwards over the span of treatments that are included in the experiment.
2. if there is a possibility of a general bending of the function in an upward or downward direction.
3. if there are any signs that indicates more complex trends.

The purpose with this method is to use polynomials which separate the effects of linear and nonlinear components that indicate the general form of relationships and approximate curves. These polynomials are of exponential or logarithmic nature. Test of trends are first and foremost conducted by two reasons, one that is based on theoretical aspects and one that is of pure empirical or descriptive nature. The first investigates if a specific trend fits a specific theory, and the other is a post hoc analysis which looks at the simplest function that will give a significant result. The four basic trends are linear, quadratic, cubic and quartic. Linear trend is the simplest form and is most often first considered. A linear trend exists if the different means from the intervals fall on the regression line and involves a theory in which a single process changes at a constant rate. A quadratic process is a single bending either upwards or downwards. Cubic and quartic trends are more complex and have two and three bendings respectively. This generalizes also to higher orders of trends, i.e. if y is of the power of p, then the curve associated with y_p has $p-1$ bendings.

4 Dummy Coding

The procedure when constructing dummy code is to first make the choice of code and then take all consideration of the interpretation of the problem from the choice of code. If you have k groups, you need to create $k-1$ vectors. In the first vector you assign group 1 a "1", and assign all other groups a "0" for that vector. In the second vector you assign the second group a "1" and all other groups a "0". This procedure is performed for all vectors. According to [5], this is the simplest code structure that allows us to investigate differences in group means, e.g. simple ANOVA.

Let us take an example with four groups.

Group	*X1*	*X2*	*X3*
Group 1	1	0	0
Group 2	0	1	0
Group 3	0	0	1
Group 4	0	0	0

According to O'Grady and Medoff in [6], dummy coding yields the same sum of squares, SS_{Adj}, as other coding techniques but only under some specific

circumstances. These are a) if the analysis does not involve any interaction terms, b) if the analysis of orthogonal design in which tests of significance is tested with a method where the variance is associated with a given predictor and adjusted for the effects in some specific subset of the other prediction variables in the equation, and c) analysis of nonorthogonal designs in which the variance is associated in the same way as in b).

Dummy coding does not proceed from any consideration of the contrasts. When there are interaction terms involved in MRA for dummy coding, special conditions must be taken into consideration. These will not be mentioned in this paper, but can studied in the paper [6].

5 Effect Coding

Effect coding is very similar to dummy coding and is created with the same process as in dummy coding with the exception that the last group will be assigned -1 for all contrasts, so only $k-1$ contrasts will be used. As in dummy coding, there is no consideration of the contrasts in effect coding. This code structure has use that is more extensive than dummy coding, but is not as fully developed as contrast coding. The interpretation is also easier to understand than the one for dummy coding. In effect coding the slope of the difference between the mean of the group is coded as "1" and the mean of all groups. One limitation is that this method only tests simple contrasts and does not allow tests of both simple and complex contrasts [2].

Let us take an example and compare it to the example for dummy coding.

Group	*X1*	*X2*	*X3*
Group 1	1	0	0
Group 2	0	1	0
Group 3	0	0	1
Group 4	-1	-1	-1

We can see that we have the same code as for dummy code, but with the exception that all elements in group 4 will have the value -1.

6 Contrast Coding

6.1 Contrast Coding

Let us start with a quote from [5];

> The choice of a coding scheme is equivalent to the choice of a set of planned ANOVA contrasts among the means of the levels of the factor.

A set of MRA codes or ANOVA contrasts are best defined as a series of precise comparisons among the group mean. Contrast coding is most appropriate when

we have a couple of hypotheses that we want to test. This is the most extensive and complicated coding method [5].

There are different approaches to design coding schemes. Serlin and Levin in [1] present the following guidelines. One is to first estimate the code values and take all considerations from that choice of code. The problem with this is that the researcher can be confused over the interpretation of the result. Another way to attack the problem is to first consider which parameters that will be estimated and then derive a set of code values that allows interpretable parameter estimates.

Let us say for example that

$$B_0 = \bar{Y}_3,$$
$$B_1 = \bar{Y}_1 - \bar{Y}_3,$$

and

$$B_2 = \bar{Y}_2 - \bar{Y}_3.$$

It can then be found for each of the two code values, X_{k1} and X_{k2}, which satisfy the equation

$$\bar{Y}_k = B_0 + B_1 X_{k1} + B_2 X_{k2}$$

that gives us the following equation

$$\bar{Y}_k = \bar{Y}_3 + (\bar{Y}_1 - \bar{Y}_3)X_{k1} + (\bar{Y}_2 - \bar{Y}_3)X_{k2}$$

and the simlpified expression is

$$\bar{Y}_k = \bar{Y}_1 X_{k1} + \bar{Y}_2 X_{k2} + \bar{Y}_3(1 - X_{k1} - X_{k2})$$

We now try to solve the equation for each value of k.

$$\bar{Y}_1 = \bar{Y}_1 X_{11} + \bar{Y}_2 X_{12} + \bar{Y}_3(1 - X_{11} - X_{12})$$

$X_{11} = 1$, $X_{12} = 0$

$$\bar{Y}_2 = \bar{Y}_1 X_{21} + \bar{Y}_2 X_{22} + \bar{Y}_3(1 - X_{21} - X_{22})$$

$X_{21} = 0$, $X_{22} = 1$

$$\bar{Y}_3 = \bar{Y}_1 X_{31} + \bar{Y}_2 X_{32} + \bar{Y}_3(1 - X_{31} - X_{32})$$

$X_{31} = 0$, $X_{32} = 0$
This will give us the codes

Group	X_{k1}	X_{k2}
Group 1	1	0
Group 2	0	1
Group 3	0	0

Another method to use is the matrix approach, where we have $\bar{Y} = (\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k)$, the vector matrix $B = (\beta_0, \beta_1, \ldots, \beta_{k-1})$ which contains the regression coefficients and the matrix X that contains the coded values that are to be solved for. This method however, requires each regression coefficient to be a linear combination of the Y means. Then the regression equation becomes

$$\bar{Y} = XB$$

If the coefficients of the desired linear combinations are represented by the matrix L, then

$$B = L\bar{Y}$$

and

$$\bar{Y} = XL\bar{Y}$$

The desired codes then satisfy the equation

$$XL = I$$

This means that $X = L^{-1}$ if all the degrees of freedom between groups are represented by contrasts. If we now that the same example as above and use the matrix approach we will get

$$L = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}$$

Since $L^{-1} = X$ we get

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Regardless of which of these methods we use, we can see that this is the same as dummy code that has been discussed in section 4. This process that we just have made is very time consuming and difficult when dummy coding already is an existing form of MRA coding. There are some different types of contrasts that are typically used in ANOVA designs that can be good to know about when constructing coding schemes; I would like to mention six types.

6.2 Simple Contrast

Simple contrasts compare means of each level of factor with the mean of a reference category. This category can be any level of factor but are usually the first or the last. An example of a simple contrast has already been seen in the section above where the contrasts are

$$B_0 = \bar{Y}_3,$$

$$B_1 = \bar{Y}_1 - \bar{Y}_3,$$

and

$$B_2 = \bar{Y}_2 - \bar{Y}_3.$$

This gives us the matrix

$$L = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}$$

and the code matrix

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

6.3 Repeated Contrast

Repeated contrasts, are like simple contrasts, pairwise comparisons among treatments. The difference between simple and repeated contrasts is that repeated contrasts compare the mean of each level to the mean of the level immediately following it. Example:

$$B_0 = \bar{Y}_1 - \bar{Y}_2,$$
$$B_1 = \bar{Y}_2 - \bar{Y}_3,$$
$$B_2 = \bar{Y}_3.$$

This gives us an L that is

$$L = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}$$

which gives us the coded matrix (with the procedure described in section 6.1)

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & -1 & 0 \\ 1 & -1 & -1 \end{pmatrix}$$

It should be noted that the code for repeated contrasts has a characteristic pattern that is easy to recognize.

6.4 Deviation Contrast

Deviation contrasts compare each individual mean to the mean of all treatments. Example:

$$B_0 = \bar{Y}_1 - \frac{1}{3}(\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3)$$

$$B_1 = \bar{Y}_2 - \frac{1}{3}(\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3)$$

$$B_2 = \bar{Y}_3 - \frac{1}{3}(\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3)$$

$\Rightarrow$

$$B_0 = \frac{2}{3}\bar{Y}_1 - \frac{1}{3}\bar{Y}_2 - \frac{1}{3}\bar{Y}_3$$

$$B_1 = -\frac{1}{3}\bar{Y}_1 + \frac{2}{3}\bar{Y}_2 - \frac{1}{3}\bar{Y}_3$$

$$B_2 = -\frac{1}{3}\bar{Y}_1 - \frac{1}{3}\bar{Y}_2 + \frac{2}{3}\bar{Y}_3$$

This satisfies the equation

$$\bar{Y}_k = \frac{2}{3}\bar{Y}_1 - \frac{1}{3}\bar{Y}_2 - \frac{1}{3}\bar{Y}_3 + (-\frac{1}{3}\bar{Y}_1 + \frac{2}{3}\bar{Y}_2 - \frac{1}{3}\bar{Y}_3)X_{k1} + (-\frac{1}{3}\bar{Y}_1 - \frac{1}{3}\bar{Y}_2 + \frac{2}{3}\bar{Y}_3)X_{k2}$$

$\Rightarrow$

$$\bar{Y}_k = (\frac{2}{3} - \frac{1}{3}X_{k1} - \frac{1}{3}X_{k2})\bar{Y}_1 + (\frac{2}{3}X_{k1} - \frac{1}{3} - \frac{1}{3}X_{k2})\bar{Y}_2 + (\frac{2}{3}X_{k2} - \frac{1}{3} - \frac{1}{3}X_{k1})\bar{Y}_3$$

Solving the equation for $k = 1, 2, 3$ yields

$$\bar{Y}_1 = (\frac{2}{3} - \frac{1}{3}X_{11} - \frac{1}{3}X_{12})\bar{Y}_1 + (\frac{2}{3}X_{11} - \frac{1}{3} - \frac{1}{3}X_{12})\bar{Y}_2 + (\frac{2}{3}X_{12} - \frac{1}{3} - \frac{1}{3}X_{11})\bar{Y}_3$$

$X_{11} = 0$, $X_{12} = -1$

$$\bar{Y}_2 = (\frac{2}{3} - \frac{1}{3}X_{21} - \frac{1}{3}X_{22})\bar{Y}_1 + (\frac{2}{3}X_{21} - \frac{1}{3} - \frac{1}{3}X_{22})\bar{Y}_2 + (\frac{2}{3}X_{22} - \frac{1}{3} - \frac{1}{3}X_{21})\bar{Y}_3$$

$X_{21} = 2$, $X_{22} = 0$

$$\bar{Y}_3 = (\frac{2}{3} - \frac{1}{3}X_{31} - \frac{1}{3}X_{32})\bar{Y}_1 + (\frac{2}{3}X_{31} - \frac{1}{3} - \frac{1}{3}X_{32})\bar{Y}_2 + (\frac{2}{3}X_{32} - \frac{1}{3} - \frac{1}{3}X_{31})\bar{Y}_3$$

$X_{31} = 0$, $X_{32} = 2$

This gives us the coded matrix

$$X = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

6.5 Helmert Contrast

Helmert contrasts compare the mean of each level with the mean of the succeeding levels. Say that we have a one-factor design with three levels, and then our contrasts will look like

$$\begin{aligned} B_0 &= \frac{1}{3}(\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3) \\ B_1 &= \bar{Y}_1 - \frac{1}{2}(\bar{Y}_2 + \bar{Y}_3) \\ B_2 &= \bar{Y}_2 - \bar{Y}_3 \end{aligned}$$

Since these codes are orthogonal the magnitude of the regression coefficients needs to be the difference among the weighted means [5].

If we use the matrix approach, we will get

$$L = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 1 & -1 \end{pmatrix}$$

and

$$X = \begin{pmatrix} 1 & \frac{2}{3} & 0 \\ 1 & -\frac{1}{3} & -\frac{1}{2} \\ 1 & -\frac{1}{3} & -\frac{1}{2} \end{pmatrix}$$

Another way to construct a matrix with Helmert contrasts is by inserting the Helmert contrasts in each column. If we need k orthogonal variables, X_1, X_2, ...,X_k, then we need a matrix with k columns and $k+1$ rows and enter the contrasts column-wise. In the first column the first element should be k and the rest of the elements in the column should be -1. Assign the first element in the second column to 0, the second element to $k-1$ and -1 to all other elements in the column. In the third column should the two first elements be 0's, the third element should be $k-2$, and all other elements in the column should be -1. Continue this process until the k:th column where all elements should be assigned 0's except for the last two elements which will be 1 and -1.

X_1	X_2	X_3	$\ldots$	X_k
k	0	0	$\ldots$	0
-1	$k-1$	0	$\ldots$	0
-1	-1	$k-2$	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
-1	-1	-1	$\ldots$	1
-1	-1	-1	$\ldots$	-1

So if we take an example with four orthogonal variables, we get

X_1	X_2	X_3	X_4
4	0	0	0
-1	3	0	0
-1	-1	2	0
-1	-1	-1	1
-1	-1	-1	-1

This method does not involve as much calculations as the first method.

6.6 Difference Contrast

Difference contrasts are also referred to as reverse Helmert contrasts since the procedure is the reversed of the Helmert contrasts. So if we as in the last section have a one-factor design with three levels, our contrasts will be

$$\begin{aligned} B_0 &= \bar{Y}_2 - \bar{Y}_1 \\ B_1 &= \bar{Y}_3 - \frac{1}{2}(\bar{Y}_1 + \bar{Y}_2) \\ B_2 &= \frac{1}{3}(\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3) \end{aligned}$$

The matrix approach gives us

$$L = \begin{pmatrix} -1 & 1 & 0 \\ -\frac{1}{2} & -\frac{1}{2} & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

and

$$X = \begin{pmatrix} \frac{1}{2} & -\frac{1}{3} & 1 \\ -\frac{1}{2} & -\frac{1}{3} & 1 \\ 0 & \frac{2}{3} & 1 \end{pmatrix}$$

7 Orthogonality and Elimination of Confounding

We start with a definition of orthogonality. Say that we have two contrasts $\sum_{i=1}^{a} c_i\mu_i$ and $\sum_{i=1}^{a} d_i\mu_i$. Then the two contrasts are orthogonal if $\sum c_i d_i = 0$.

Example:

Treatment	c_i	d_i
1	-2	0
2	1	-1
3	1	1

Since $-2 \cdot 0 + (-1) \cdot 1 + 1 \cdot 1 = 0$ the contrasts are orthogonal. Each column should also sum to zero.

Orthogonal contrasts are uncorrelated variables created to test uncorrelated hypotheses. O'Grady and Medoff [6] defines orthogonal design as "any multi-factor design in which the frequency of observations within each cell is equal". Nonorthogonal design then of course is any multi-factor design which the frequency of observations within each cell is unequal, but also disproportionate.

The order of elimination of confounding is a method in which the effect of interest is ignored and/or eliminated from each other in the analysis of a nonorthogonal design. It is necessary to eliminate some degrees of confounding in a nonorthogonal design since some or all of the effects are correlated with each other. Four methods for elimination of confounding are mentioned.

1. In this method each effect is adjusted for every other effect in the model. This method is referred to as the "regression approach" or "full rank approach".

2. The second method means adjustment of each effect at a given degree of complexity for all other effects of that degree and all effects of lower degree. This method is referred to as "traditional ANOVA".

3. The third method evaluates a single model in which the effects is entered in a specific order. Each effect is adjusted for all previous effects in the model. Effects of higher order cannot precede effects of lower order in the model. This means that the interaction term cannot be evaluated and ignores the involvement of the main effect in the interaction term. This is called a "hierarchical approach".

4. This last method involves when ignoring all other effects of the same or more complex order, each effect is adjusted for all simpler effects. This method is called "weighted means approach".

In an orthogonal design will all these methods of elimination of confounding give the exact same sum of squares and test of significance for a specific effect. In a nonorthogonal design will none of these methods give the same sum of squares (if not by chance) since each method of elimination of confounding means different model comparisons. Say that we have the following models

$$1)\ y_{ijk} = \mu + \varepsilon_{ijk}$$
$$2)\ y_{ijk} = \mu + \alpha_j + \varepsilon_{ijk}$$
$$3)\ y_{ijk} = \mu + \beta_k + \varepsilon_{ijk}$$
$$4)\ y_{ijk} = \mu + \alpha_j + \beta_k + \varepsilon_{ijk}$$
$$5)\ y_{ijk} = \mu + \alpha_j + (\alpha\beta)_{jk} + \varepsilon_{ijk}$$
$$6)\ y_{ijk} = \mu + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk}$$
$$7)\ y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk}$$

Let us now compare these models for different effects

Comparison	*A*	*B*	*AB*
1	7-6	7-5	7-4
2	4-3	4-2	7-4
3	2-1	4-2	7-4
4	2-1	3-1	7-4

As can be seen in the tables above, the effects A and B will answer different questions if the design is nonorthogonal. Comparison (1) compares model 7 and 6 for the main effect α_j, 7 and 5 for main effect β_k and 7 and 4 for the interaction term $(\alpha\beta)_{jk}$. We can see that comparison (1) answers the question if each main effect is considered as a unique variance in Y if that effect was added last in the equation, (2) accounts for if the main effect is considered as a significant amount of variance in Y over the expected for the second main effect when the interaction term is ignored, (3) answers the question if the first main effect in the equation is considered as a significant amount of variance in Y, when ignoring all other main effects and interaction terms. It also answers the question if the second main effect that is added to the equation is considered as a significant amount of variance in Y over the accounted for by the other main effect when ignoring the interaction term, and (4) states if each main effect in the model can be considered as a significant amount of Y if all other effects in the model is ignored [6].

As can be seen there are a lot of things to take into consideration when working with nonorthogonal contrasts, which is why orthogonal contrasts are preferred.

8 Implementations in R

8.1 Introduction to contrast coding in R

All of the coding techniques that have been mentioned in Section 6 can be performed in the statistical program R. We can begin with repeating the fact that no matter which coding technique that are used, if we have a categorical variable with k levels, then we will have a contrast matrix with k rows and $k-1$ columns. The first thing that needs to be done is to read the data frame and create a factor variable.

Say that we have a data file called `file.csv` and we want to create a factor variable called `factorvar`. The file format csv stands for *comma separated values*. We use the commands

```
file = read.table('c:/file.csv', header=T, sep=",")
# creating the factor variable factorvar
factorvar = factor(variable, labels=c("A", "B", "C", "D"))
```

(Note that this is a general example and the file does not exist.)

8.2 Simple Contrast

As mentioned in Section 6.2, simple contrasts compare each level of factor with the mean of a reference category. There are two ways to create a simple contrast matrix in R; the first way is to use the built-in contrast function `contr.treatment` and the other is to manually create the contrast matrix.

The built-in function `contr.treatment` uses group 1 as a reference group and the command must be specified for how many levels that are included. The command is `contr.treatment(k)` for `k` levels. So if we have for example $k = 4$ levels, we will have the command `contr.treatment(4)`.
This gives the result in R

	2	*3*	*4*
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

You now assign the treatment contrast to `factorvar` with the following command

```
contrasts(file$factorvar) = contr.treatment(4)
# And the regression with the usual lm command
summary(lm(dependentvariable ~ factorvar, file))
```

If we are to perform simple contrast coding manually we first need to create a contrast scheme. In the general case with k levels we have the scheme

Level				
1	$-\frac{1}{k}$	$-\frac{1}{k}$	$\dots$	$-\frac{1}{k}$
2	$\frac{(k-1)}{k}$	$-\frac{1}{k}$	$\dots$	$-\frac{1}{k}$
3	$-\frac{1}{k}$	$\frac{(k-1)}{k}$	$\dots$	$-\frac{1}{k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$-\frac{1}{k}$	$-\frac{1}{k}$	$\dots$	$\frac{(k-1)}{k}$

So if we have $k = 4$ levels, we will get the coding scheme

Level			
1	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$
2	$\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$
3	$-\frac{1}{4}$	$\frac{3}{4}$	$-\frac{1}{4}$
4	$-\frac{1}{4}$	$-\frac{1}{4}$	$\frac{3}{4}$

You then create the matrix manually in R, let us call the matrix `my.simple`

```
my.simple = matrix(c( -1/4, 3/4, -1/4, -1/4,
-1/4, -1/4, 3/4, -1/4,  -1/4, -1/4, -1/4, 3/4), ncol=3)
contrasts(file$factorvar) = my.simple
```

We can now use the `lm` command as usual

```
summary(lm(dependentvariable~factorvar, file))
```

8.3 Deviation Contrast

As can be studied in Section 6.4, deviation contrasts compare each individual mean to the mean of all treatments. There is no need to construct a contrast matrix since the built-in function `contr.sum` does that for us. If we have k levels, the command will be `contr.sum(k)`. If we as in the example above have $k = 4$ levels, our code will be `contrasts(file$factorvar) = contr.sum(4)`. Regression is made with the usual `lm` command.

8.4 Helmert Contrast

As has been seen in Section 6.5, Helmert contrasts compare the mean of each level with the mean of the succeeding levels.

The general contrast matrix will look like

$$\begin{array}{cccccc} \textit{Level} & & & & \\ 1 & \frac{(k-1)}{k} & 0 & \dots & 0 \\ 2 & -\frac{1}{k} & \frac{(k-2)}{k-1} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ k-1 & -\frac{1}{k} & -\frac{1}{k-1} & \dots & \frac{1}{2} \\ k & -\frac{1}{k} & -\frac{1}{k-1} & \dots & -\frac{1}{2} \end{array}$$

If we have as in the last example $k = 4$ levels the matrix will be

$$\begin{array}{cccc} \textit{Level} & & & \\ 1 & \frac{3}{4} & 0 & 0 \\ 2 & -\frac{1}{4} & \frac{2}{3} & 0 \\ 3 & -\frac{1}{4} & -\frac{1}{3} & \frac{1}{2} \\ 4 & -\frac{1}{4} & -\frac{1}{3} & -\frac{1}{2} \end{array}$$

To create the matrix in R, we use the command

```
my.helmert = matrix(c(3/4, -1/4, -1/4, -1/4, 0, 2/3,
-1/3, -1/3, 0, 0, 1/2, -1/2), ncol = 3)
# and we assign the new Helmert code to factorvar
contrasts(file$factorvar) = my.helmert
```

Again, the `lm` command is used to do regression. There is a built-in function `contr.helmert`, but the method is discussed in [7] whether it is good to use or not, and it is therefore better to do this method manually.

8.5 Difference Contrast

Difference contrast is also referred to as reverse Helmert contrast (see Section 6.6), which compares each level with the mean of the previous level(s). The general procedure is the reversed of the one in Section 8.4. If we have $k = 4$ levels, the matrix will be

Level			
1	$\frac{1}{2}$	$-\frac{1}{3}$	$-\frac{1}{4}$
2	$-\frac{1}{2}$	$-\frac{1}{3}$	$-\frac{1}{4}$
3	0	$\frac{2}{3}$	$-\frac{1}{4}$
4	0	0	$\frac{3}{4}$

The R code to create this matrix is

```
my.difference = matrix(c(-1/2, 1/2, 0, 0, -1/3, -1/3, 2/3, 0,
 -1/4, -1/4, -1/4, 3/4), ncol = 3)
# We assign the difference contrast to factorvar
contrasts(file$factorvar) = my.difference
```

Regression is made with the usual `lm` command.

8.6 User defined Contrast

It is possible to use any kind of coding scheme. Let us say for example that we want to make the following comparisons

Level 1 to level 3: $B_0 = \bar{Y}_1 - \bar{Y}_3$
Level 2 to levels 1 and 4: $B_1 = \bar{Y}_2 - \frac{1}{2}(\bar{Y}_1 + \bar{Y}_4)$
Levels 1 and 2 to levels 3 and 4: $B_2 = \bar{Y}_1 + \bar{Y}_2 - (\bar{Y}_3 + \bar{Y}_4)$

The code in R will then be

```
mat = matrix(c(1, 0, -1, 0, -1/2, 1, 0, -1/2, -1/2, -1/2, 1/2,
1/2), ncol = 3)
# To create the contrasts that we are interested of use the command
my.contrasts = mat \% * \% solve(t(mat) \% * \%  mat)
# Assigning my.contrasts to factorvar
contrasts(file$factorvar) = my.contrasts
```

9 Conclusion

In this article coding schemes and coding techniques have been discussed. One useful aspect with coding schemes is that you can change qualitative data into quantitative data to make mathematical calculations possible. Another issue is that you can take large amounts of data that normally would take a lot of

time to calculate and transform it into 1's and 0's to make the calculations more calculating effective. To be able to create a coding scheme it is important to have a clear vision of what questions are to be answered. If the researcher does not have a clear vision of what he wants to investigate it might be difficult to choose the coding technique that is best suited in that specific case. If the researcher does not have a clear vision of the problem the result might also be difficult to interpret. In this paper have also the possibilities with MRA been pointed out. To understand and construct different types of coding schemes, the use of MRA is a helpful tool, which has been demonstrated in literature.

References

[1] Ronald C. Serlin and Joel R. Levin. *Teaching How to Derive Directly Interpretable Coding Schemes for Multiple Regression Analysis.* Journal of Educational and Behavioral Statistics 10:223 (1985).

[2] Matthew J. Davis. *Contrast Coding in Multiple Regression Analysis:Strengths, Weaknesses, and Utility of Popular Coding Structures.* Journal of Data Science 8 (2010).

[3] Kathy Chatham. *Planned Contrsasts: An Overview of Comparision Methods.* Paper presented at the Annual Meeting of the Soutwest Educational Research Association (San Antonio, TX, January, 1999)

[4] Wilda Laija. *Conducting ANOVA Trend Analyses Using Polynomial Contrasts.* Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January, 1997)

[5] Craig A. Wendorf. *Primer on Multiple Regression Coding: Common Forms and the Additional Case of Repeated Contrasts.* Understanding Statistics, 3(1) (2004).

[6] Kevin E. O'Grady and Deborah R. Medoff. *Categorical Variables in Multiple Regression: Some Cautions.* Multivariate Behavioral Research, 23 (1988).

[7] W. N. Venables and B.D. Ripley. *Modern Applied Statistics with S.* Springer-Verlag (2002).