Hutcheson, G. D. (2011). Categorical Explanatory Variables. Journal of Modelling in Management. 6, 2: 225–236.
http://www.emeraldinsight.com/loi/jm2
http://dx.doi.org/10.1108/jm2.2011.29706baa.002

Journal of Modelling in Management

Graeme D. Hutcheson

Categorical explanatory variables

NOTE: this is a slightly updated version of this paper which is distributed to correct some formatting difficulties with the web version.

Many models require the addition of categorical data as explanatory variables. Although the techniques to do this are well-known, it is common to only see the default methods used or categorical variables incorrectly analysed as numeric (particularly if the software used utilises a numeric coding scheme for categorical data; see Hutcheson, 2011a). This tutorial outlines some options for analysing categorical data and provides some example analyses showing the advantages of using different coding schemes.

Categorical variables (ordered and unordered) are very common in social science research and are often of primary interest. In order to include these variables appropriately into statistical models they need to be coded into a number of individual "dummy" categories, which can be entered directly into the model. There are a variety of methods that can be used to code these dummy-variables, each of which provides a different set of comparisons between the categories that make up the variable. Some coding techniques compare individual categories, others compare specific categories with mean values whilst others provide information about possible linear and non-linear trends. These coding methods provide a wealth of information that can be of great benefit to researchers.

Even though many different coding methods are available for categorical data, researchers tend to opt for the simplest method of coding, or use the default method offered by their statistical software. The default or simplest coding is often not, however, the most appropriate or useful way to represent the categorical variable, particularly when the variable is ordered, or specific comparisons are required.

This tutorial provides a demonstration of a number of methods for coding categorical explanatory variables and shows how these can be used to describe ordered and well as unordered categories. The use of these coding methods can greatly improve the interpretation of the results and enhance analyses.

1. Coding Unordered Data:

The following example data shows the price of a “standard drink” and the location of the bar where the drink was purchased. Three different locations are shown; the town centre, the sea front and other areas. Figure 1 displays these data in a box-plot and clearly shows that sea-front bars tend to charge the most, closely followed by bars in the town centre with those in other locations charging somewhat lower prices.

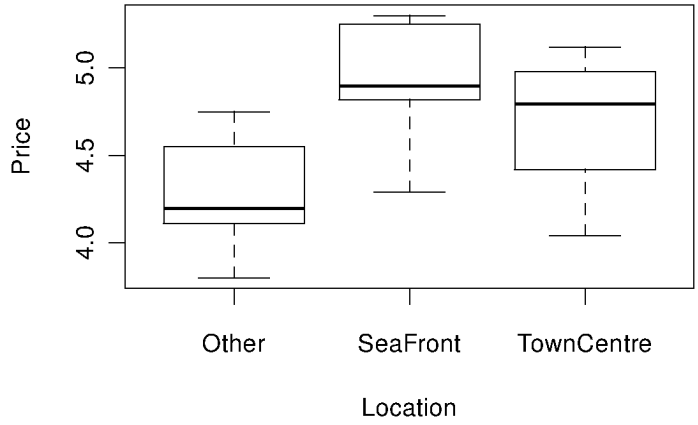

Figure 1: The relationship between Price and Location.

The relationship between Price and Location can also be described using an OLS regression model of “Price” with “Location” included as an explanatory variable. In order to include the categorical variable “Location” in the regression model, it is necessary to dummy-code it (either by hand or by software). Below are shown two popular methods of including unordered categorical variables in statistical models.

1.1 Treatment Coding (comparing each category to a reference)

One of the most popular methods for coding categorical data is a technique known as treatment coding (also known as indicator or simple coding) which transforms a categorical variable into a number of dichotomies. Table 1 shows how the variable Location may be coded into a series of dichotomies.

Table 1: Treatment Coding of Location

		Dummy Codes		
	Location	**D.Other**	**D.SeaFront**	**D.TownCentre**
Categories	**Other**	1	0	0
	SeaFront	0	1	0
	TownCentre	0	0	1

The relationship between Price and Location can be investigated using a regression model that substitutes the dummy codes for the original variable. In general, if we have *j* categories, *j*-1 dummy variables may be entered into the model. The three-category variable “Location” is, therefore, represented using two dummy variables, each of which indicates a specific location that is compared

to the reference category (the location that is not included as a parameter). Although many software packages dummy-code automatically "in the background", dummy codes can also be entered directly into the data-frame (the spreadsheet containing the data). The data-frame in Table 2 shows the variable "Location" and the treatment-coded dummy variables (D.Other is missing, as this is the reference category). We can either model "Price" using the variable "Location" (if our software allows automatic dummy coding of categorical data), or model "Price" using both the dummy-variables "D.SeaFront" and "D.TownCentre". The resulting models will be identical (try it and see!).

Table 2: A data-frame showing Treatment Coding of Location (reference category="other")

Price	Location	D.SeaFront	D.TownCentre
5.43	TownCentre	0	1
5.02	TownCentre	0	1
4.76	Other	0	0
6.73	SeaFront	1	0
4.98	Other	0	0
5.32	TownCentre	0	1
5.72	SeaFront	1	0
5.47	SeaFront	1	0

Running the regression model: the default option

A regression model of Price (an OLS model; Price ~ Location) computed in R (2011) via the Rcmdr interface (Fox, 2011), provides the following output:

```
OLS regression Model:  Price ~ Location
                       Treatment contrasts for Location (ref = Other)

                       Estimate  Std. Error  t value  Pr(>|t|)
Location[T.SeaFront]     0.6140      0.1446    4.246  0.000230 ***
Location[T.TownCentre]   0.4170      0.1446    2.883  0.007631 **

F-statistic: 9.398 on 2 and 27 DF,  p-value: 0.0007983
```

Although Location is a single variable, it is represented in the model as two (*j*-1) dummy variables. The default coding used by R is treatment coding (hence the letter "T" in the parameter description) with the reference category being the first category alphabetically (the category "Other"). The first Location parameter compares "SeaFront" with "Other" and the second compares "TownCentre" with "Other". The software has simply re-coded "Location" in the background (without saving these codes to the dataset).

In R, it is simple to show the contrasts used in the regression model using the "`contrasts()`" command. For example, to show which contrasts are being used for the variable "Location" (which is contained in the data-frame "BarPrices") the following command...

```
> contrasts(BarPrices$Location)

           [T.SeaFront] [T.TownCentre]
Other                 0              0
SeaFront              1              0
TownCentre            0              1
```

shows that treatment codes are used (indicated by “T.”) and “Other” is the reference category as it is the dummy code that is missing. It is important to note that this default coding does not provide a complete picture of the relationship between Price and Location. One obvious difficulty with the model above is that it does not allow us to directly compare all locations. We have a parameter that compares SeaFront with Other (the `Location[T.SeaFront]` parameter) and one that compares TownCentre with Other (the `Location[T.TownCentre]` parameter), but not one that compares SeaFront with TownCentre. To do this, we need to change the reference category.

Changing the reference category:

The reference category for the variable Location (contained within the BarPrices dataset) can be changed to TownCentre easily in Rcmdr using pull-down menus, or directly In R using the command...

```
> BarPrices$Location <- factor(BarPrices$Location,
                          levels=c('TownCentre','SeaFront','Other'))
```

and checked using...

```
> contrasts(BarPrices$Location)

             [T.SeaFront] [T.Other]

TownCentre              0         0
SeaFront                1         0
Other                   0         1
```

Dummy categories are now provided for “SeaFront” and “Other”, making “TownCentre” the reference category. Changing the reference category is usually very simple to do using most software - refer to the relevant manual for instructions. Changing the reference category to TownCentre produces the following model (an OLS model; Price ~ Location) :

```
OLS regression Model:  Price ~ Location
                        Treatment contrasts for Location (ref = TownCentre)

                      Estimate  Std. Error  t value  Pr(>|t|)
Location[T.SeaFront]    0.1970      0.1446    1.362   0.18439
Location[T.Other]      -0.4170      0.1446   -2.883   0.00763 **

F-statistic: 9.398 on 2 and 27 DF,  p-value: 0.0007983
```

Changing the reference category looks to have made a huge difference to the parameters for bars located on the SeaFront. In the first model, it is highly significant and in the second, it is non-significant. Although this is to be expected given the different comparisons being made in the model (a quick look at Figure 1 will confirm that the difference between SeaFront and Other is large compared to the difference between SeaFront and TownCentre), it can be misleading if just one model is shown, particularly to audiences not used to dummy-coded explanatory variables. The first model makes Location look much more significant than the second! (when in fact, both models are identical).

Showing all comparisons:

If the explanatory variable is of particular interest, it is often useful to construct a table showing all comparisons (these have been compiled from information gathered using the two models above). Table 3 shows the individual comparisons for the model Price ~ Location.

Table 3: A table of comparisons for Location. Each category is compared to a reference category. The values show the difference in price between the categories and the stars indicate significance. For example, TownCentre bars are 0.197 cheaper than SeaFront bars, a difference that is not significant.

		Compared to...		
		Other	**SeaFront**	**TownCentre**
Categories	**Other**	-	-0.614 ***	-0.417 **
	SeaFront	0.614 ***	-	0.197
	TownCentre	0.417 **	-0.197	-

The significance of Location:

The models above do not provide direct information about the overall significance of the variable "Location" on "Price". In order to do this, the effect that both parameters have on Price simultaneously needs to be assessed. Although this is simply achieved in most statistical software, it is often missing in research reports and papers. It is not uncommon for readers to have to come to their own conclusions about significance based on the individual estimates of significance given in the reported model, which, as we have seen, can provide very different impressions of significance. The overall significance of Location computed using R, is shown below:.

```
Anova Table (Type II tests)

Response: Price
          Sum Sq Df F value    Pr(>F)
Location  1.9656  2  9.3983 0.0007983 ***
Residuals 2.8235 27
```

The significance of the Location variable is 0.0007983.

1.2 Sum Coding (comparing each category to the average)

It is sometimes appropriate to compare each category with an average value from all categories, rather than a specific reference. This is possible using a different dummy coding technique, where the codes are assigned according to the scheme laid out in Table 4.

Table 4: Sum Coding of Location

		Dummy Codes		
		D.Other	**D.SeaFront**	**D.TownCentre**
Categories	**Other**	1	0	0
	SeaFront	0	1	0
	TownCentre	-1	-1	-1

Using these codes, each category is compared to the average of all categories. Similar to the treatment coding method discussed above, only *j*-1 categories may enter into a model. It is (usually) a simple matter to change the coding technique used for a variable. Rcmdr uses pull-down menus to change the contrast coding method (see Figure 3), but this can also be achieved directly in R using the command...

```
> contrasts(BarPrices$Location) <- "contr.Sum"
```

and checked using the contrasts() command...

```
1> contrasts(BarPrices$Location)

           [S.Other] [S.SeaFront]
Other              1            0
SeaFront           0            1
TownCentre        -1           -1
```

which shows that sum coding is used (as indicated by S.) with TownCentre as the reference category.

Running the regression model: the default option

A regression model of Price (an OLS model; Price ~ Location) computed in R using sum coding provides the following output:

```
OLS regression Model:  Price ~ Location
                       Treatment contrasts for Location (ref = TownCentre)

                       Estimate  Std. Error  t value  Pr(>|t|)
Location[S.Other]    -0.34367    0.08350  -4.116 0.000325 ***
Location[S.SeaFront]  0.27033    0.08350   3.238 0.003184 **

F-statistic: 9.398 on 2 and 27 DF,  p-value: 0.0007983
```

The statistics for the overall model are the same as before (see the F value). The sea front bars charge significantly more than the average of all bars. To compare “TownCentre” bars to the average of all bars, the reference category can be changed (see the instructions above) and the model re-run.

```
OLS regression Model:  Price ~ Location
                       Treatment contrasts for Location (ref = SeaFront)

                       Estimate  Std. Error  t value  Pr(>|t|)
Location[S.TownCentre]  0.07333    0.08350   0.878 0.387539
Location[S.Other]      -0.34367    0.08350  -4.116 0.000325 ***

F-statistic: 9.398 on 2 and 27 DF,  p-value: 0.0007983
```

The estimate for Other is the same as before (it is still being compared to the overall average). We can see that the outlying bars charge significantly less than the average. These results are summarised below:

Table 5: A table of comparisons for Location

		Compared to the average...
Categories	**SeaFront**	**0.27033 ****
	TownCentre	**0.07333**
	Other	**-0.34367 *****

2. Coding Ordered data.

Ordered categorical explanatory variables are very common and may be used to indicate information such as educational grade, socio-economic status, attitude, experience, management level etc. The following example data shows an ordered variable (called “Variable”) with five levels and a box-plot showing its relationship to a numeric variable (called “Score”) is shown in Figure 2. Although this variable can be included as an explanatory in a model using one of the dummy-variable coding techniques described above (treatment or sum coding), these methods do not take into account the order in the data. A number of alternative coding methods are explored below that take account of order and offer advantages when analysing ordered categorical explanatory variables.

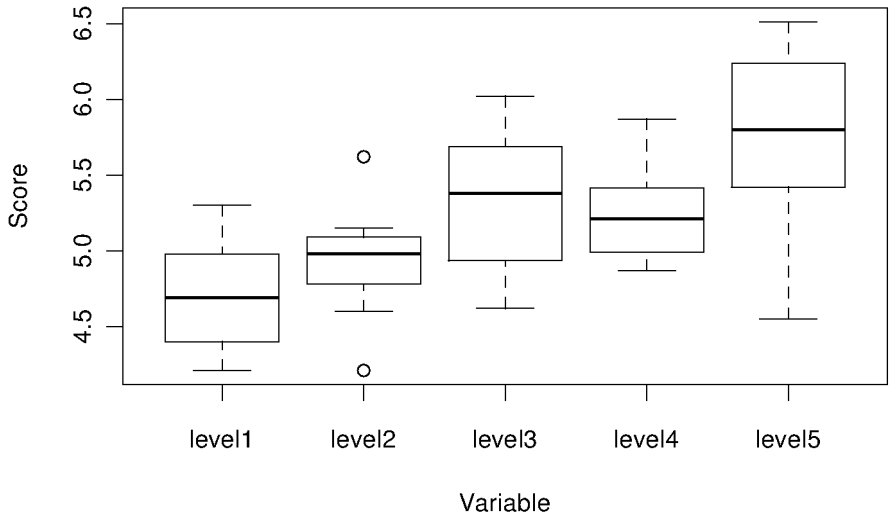

Figure 2: The relationship between Score and an ordered Variable.

2.1 Helmert Coding (comparing each level to the mean of previous levels)

One method of taking account of order in the data is to use Helmert coding, which compares individual levels to the average of previous levels. Table 6 shows the coding method used to obtain Helmert contrasts.

Table 6: Helmert Coding of Ordered variable

		Dum1	Dum2	Dum3	Dum4
levels	**Level1**	-1	-1	-1	-1
	Level2	1	-1	-1	-1
	Level3	0	2	-1	-1
	Level4	0	0	3	-1
	Level5	0	0	0	4

Helmert coding is defined in R using the commands...

```
> contrasts(Dataset$Variable) <- "contr.Helmert"
```

and checked using the command...

```
1> contrasts(Dataset$Variable)

       [H.1] [H.2] [H.3] [H.4]
level1    -1    -1    -1    -1
level2     1    -1    -1    -1
level3     0     2    -1    -1
level4     0     0     3    -1
level5     0     0     0     4
```

Using the Helmert contrasts for the OLS regression model “Score ~ Variable” gives the following output:

```
OLS regression Model:  Score ~ Variable
                       Helmert contrasts for Variable

            Estimate  Std. Error  t value  Pr(>|t|)
Variable1    0.11513     0.11082    1.039  0.304390
Variable2    0.17496     0.05905    2.963  0.004857 **
Variable3    0.06884     0.04084    1.686  0.098765 .
Variable4    0.13852     0.03254    4.257  0.000104 ***

F-statistic: 7.806 on 4 and 45 DF,  p-value: 7.237e-05
```

Variable1 [H.1] compares level2 with level1. We can see that level2 has a higher score than level1, but not significantly so. Variable2 [H.2] compares level3 with the average of levels 1 and 2. Variable3 [H.3] compares level4 with the average of the first 3 levels and Variable4 [H.4] compares level5 with the average of the preceding 4 levels. These parameters show an increasing trend in the data (all estimates are positive and show that each category is bigger than the average of the preceding categories). Similar to the previous coding schemes, the overall significance of the explanatory variable, cannot be assessed directly from the output – an overall test of all 4 parameters is needed (an analysis of deviance table could be used, but as there is a single explanatory variable, we will just use the overall F-test, which shows a significance of 7.237e-05).

Difference Coding (comparing each level to it's neighbour)

A useful thing to do with ordered data is to compare each level with it's neighbour. This provides information about the trend in the variable and quickly identifies levels that do not “follow the trend”. Difference coding is not one of the techniques that is automatically available in R and Rcmdr, but it can easily be implemented by specifying the contrasts manually. The procedure for achieving this in Rcmdr is shown in Figure 3 (for other software packages, please consult the manual). The coding used for each category is shown in the “Specify Contrasts” window.

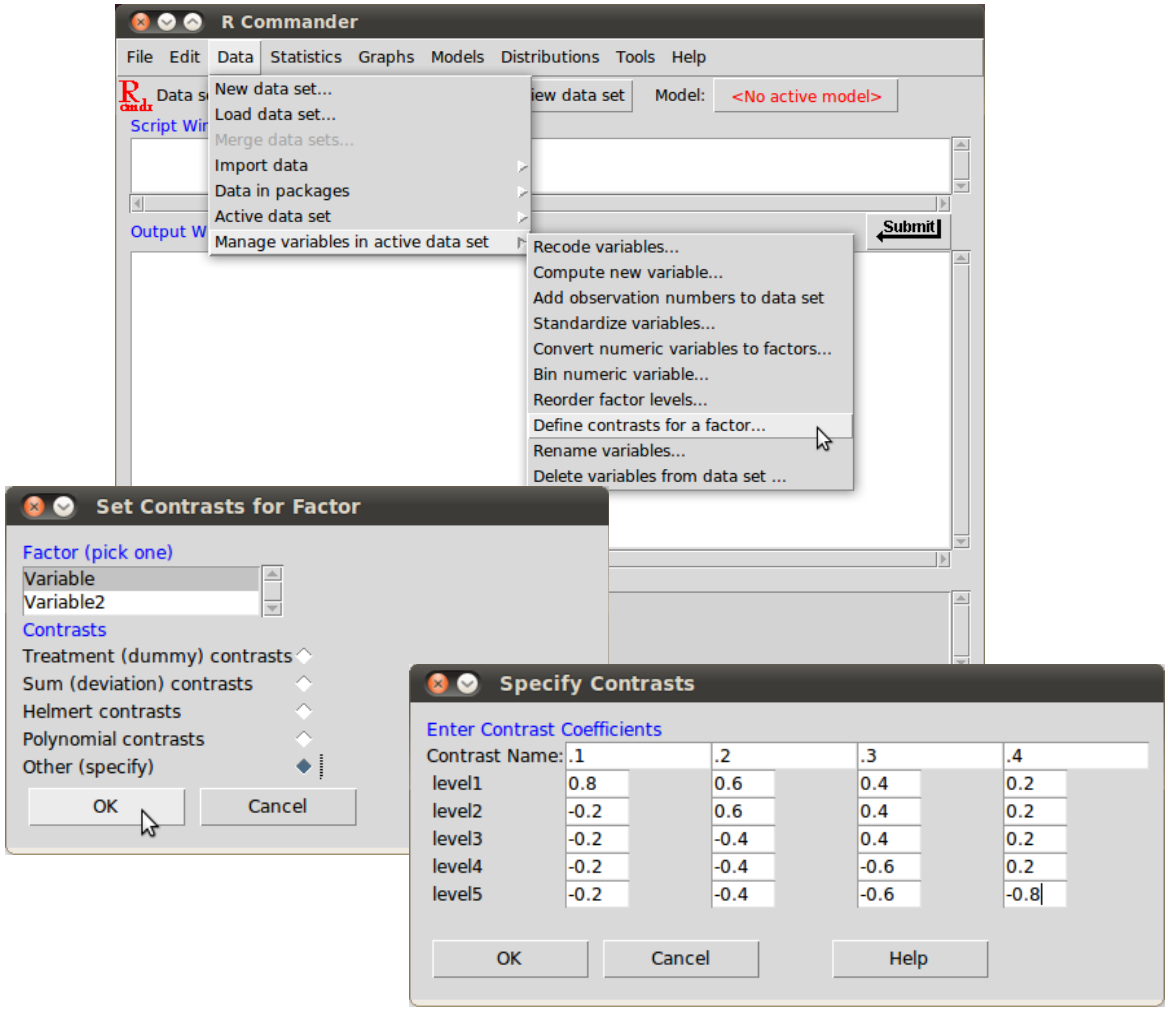

Figure 3: Difference coding in Rcmdr.

The model of “Score ~ Variable”, when using the difference coding technique is shown below:

```
OLS regression Model:  Score ~ Variable
                       Helmert contrasts for Variable

             Estimate  Std. Error  t value  Pr(>|t|)
Variable.1  -0.23026     0.22163   -1.039    0.3044
Variable.2  -0.40974     0.22163   -1.849    0.0711 .
Variable.3   0.07455     0.19546    0.381    0.7047
Variable.4  -0.48609     0.20029   -2.427    0.0193 *
F-statistic: 7.806 on 4 and 45 DF,  p-value: 7.237e-05
```

The Variable.1 parameter compares Level1 with Level2, Variable.2 compares Level2 with Level3, etc. From these parameters it is immediately obvious that Levels 3 and 4 do not follow the same pattern as the others as it has a positive parameter estimate (this is also evident in Figure 2, as Level4 is below level3). On the basis of this evidence, one might want to look more closely at levels 3 and 4 to see if they might be combined.

Orthogonal Polynomial Coding (identifying linear and non-linear trends)

Polynomial coding is one of the more rarely used coding techniques, but it also one of the most informative. The purpose of polynomial coding is to try and identify linear and non-linear trends in the relationship between the ordered explanatory variable and the response. This coding should only be used where the categories can be considered to be 'more or less' equally-spaced. The polynomial coding scheme for the data is shown in Table 7.

Table 7: Polynomial Coding of Ordered variable

		.L	.Q	.C	^4
levels	**Level1**	-0.632	0.535	-0.316	0.120
	Level2	-0.316	-0.267	0.632	-0.478
	Level3	0	-0.535	0	0.717
	Level4	0.316	-0.267	-0.632	-0.478
	Level5	0.632	0.535	0.316	0.120

Orthogonal polynomial coding is defined in R using the commands...

```
> contrasts(Dataset$Variable) <- "contr.poly"
```

and checked using the command...

```
> contrasts(Dataset$Variable)

                .L         .Q            .C         ^4
[1,] -6.324555e-01  0.5345225 -3.162278e-01  0.1195229
[2,] -3.162278e-01 -0.2672612  6.324555e-01 -0.4780914
[3,] -3.287978e-17 -0.5345225  2.164914e-16  0.7171372
[4,]  3.162278e-01 -0.2672612 -6.324555e-01 -0.4780914
[5,]  6.324555e-01  0.5345225  3.162278e-01  0.1195229
```

The model of “Score ~ Variable”, when using the polynomial coding technique is shown below:

```
OLS regression Model:  Score ~ Variable
                       Polynomial contrasts for Variable

             Estimate   Std. Error   t value    Pr(>|t|)
Variable.L  0.771054    0.144770    5.326  3.08e-06 ***
Variable.Q  0.007317    0.142927    0.051     0.959
Variable.C  0.120532    0.153818    0.784     0.437
Variable^4  0.204227    0.147054    1.389     0.172

F-statistic: 7.806 on 4 and 45 DF,  p-value: 7.237e-05
```

The first parameter, Variable.L, tests a linear trend, the second parameter (Variable.Q) tests for a quadratic trend (a curve), the third (Variable.C) a cubic trend. Further parameters test for higher order trends. The model shows that the relationship between the Score and the ordered variable is linear, which can be confirmed from the boxplot in Figure 2.

Ploynomial coding is particularly useful for identifying curvilinear relationships, as in the following example where successive increases in level have a decreasing effect on the response.

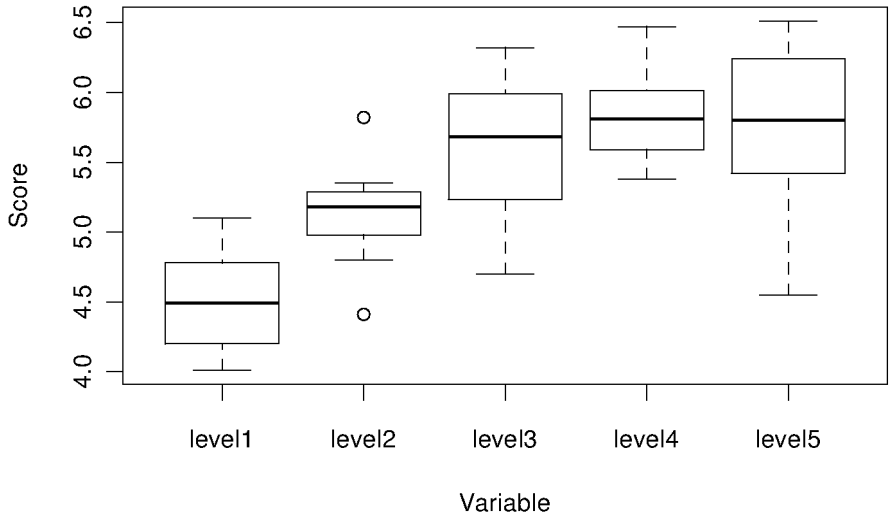

Figure 4: A curvi-linear relationship.

```
OLS regression Model:  Score ~ Variable
                Polynomial contrasts for Variable

      Estimate   Std. Error   t value  Pr(>|t|)
Variable.L    1.018286     0.149454     6.813  1.93e-08 ***
Variable.Q   -0.454316     0.147552    -3.079   0.00353 **
Variable.C   -0.057705     0.158795    -0.363   0.71801
Variable^4    0.002125     0.151813     0.014   0.98889

F-statistic: 14.88 on 4 and 45 DF,  p-value: 8.019e-08
```

This model shows a curvilinear trend, as parameters Variable.L and Variable.Q are both significant. This is precisely what one would expect from the shape of the relationship shown in Figure 4. It is also evident from the parameter Variable.Q that the quadratic effect decreases as level increases.

Polynomial contrasts are also useful for identifying non-linear trends that are difficult to identify from the regression parameters and fit statistics. For example, the relationship shown in Figure 5 does not show a linear relationship, but a quadratic one might be more useful in describing the relationship.

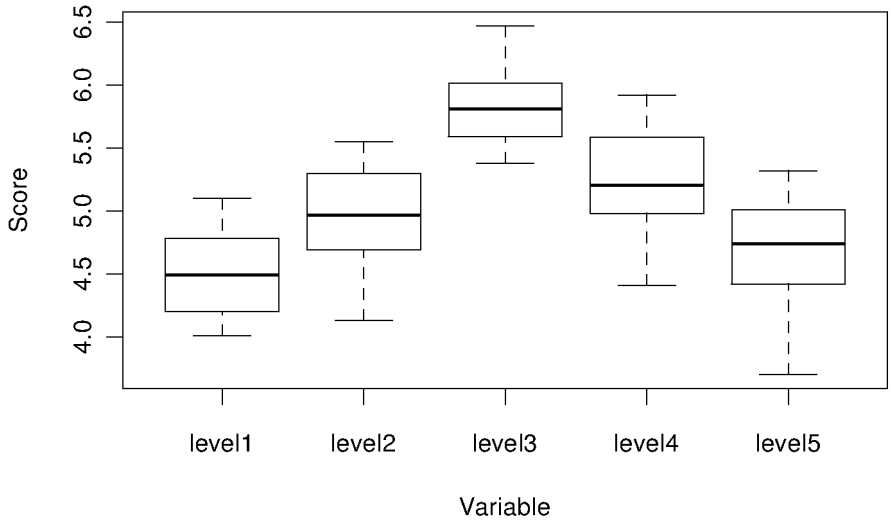

Figure 5: A non-linear relationship.

```
OLS regression Model:  Score ~ Variable
                Polynomial contrasts for Variable

      Estimate   Std. Error   t value  Pr(>|t|)
Variable.L   0.20297    0.13945   1.455  0.15248
Variable.Q  -0.93790    0.13784  -6.804 1.99e-08 ***
Variable.C  -0.11731    0.14649  -0.801  0.42746
Variable^4   0.42393    0.14089   3.009  0.00428 **

F-statistic: 15.27 on 4 and 45 DF,  p-value: 5.818e-08
```

3. Conclusion

Dummy variable coding is an important part of data manipulation as it enables categorical variables to be included in a wide variety of statistical models (for example, OLS, proportional-odds, survival, multinomial and log-linear). It's use increases the utility of regression models and understanding how the coding operates greatly helps with the interpretation of the models. Careful selection of a contrast code and a reference category is crucial to effective data analysis.

Further Reading

Aguinis, H. (2004). *Regression Analysis for Categorical Moderators*. Guilford Press.

Fox, J. and Weisberg, S. (2011). *An R and S-Plus Companion to Applied Regression (2nd edition)*. London: Sage Publications.

Hardy, M. A. (1993). *Regression with dummy variables.* London: Sage Publications

Hutcheson, G. D. (2011a). Data coding, management and manipulation. Journal of Modelling in Management, 6: 1.

Hutcheson, G. D. (2011b). Dummy Variable Coding. In L. Moutinho, L. and Hutcheson, G. D. *The SAGE Dictionary of Quantitative Management Research.* Sage Publications.

Hutcheson, G. D. and Moutinho, L. (2008). Statistical Modelling for Management. London: Sage Publications.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org

Graeme Hutcheson
Manchester University