Contrast Coding in R: An Exploration of a Dataset

Rachel Baker

Phonatics, Sept. 29, 2009

Thanks to:
http://www.ats.ucla.edu/stat/R/library/contrast_coding.htm
Roger Levy

Coding for Regressions

- Categorical variables need to be recoded into a series of variables which can then be entered into the regression model
- There are a variety of coding systems that can be used when coding categorical variables
- You should choose a coding system that reflects the comparisons that are most meaningful for testing your hypotheses

Coding for Regressions

- Coded comparisons represent planned comparisons and not post hoc comparisons
  - They are comparisons that you plan to do before you begin analyzing your data, not comparisons that you think of once you have seen the results of preliminary analyses.
- Some forms of coding make more sense with ordinal categorical variables than with nominal categorical variables

Some Coding Schemes

Contrast	**Comparison**
Simple/Treatment	Each level to reference level
Deviation	Deviations from the grand mean
Helmert	Levels of a variable with the mean of the subsequent levels
Reverse Helmert	Levels of a variable with the mean of the previous levels
Forward Difference	Each level minus the next level
Backward Difference	Each level minus the previous level
User-Defined/ Contrast	User-defined contrasts

Specifying Multiple Contrasts

- Contrast coding can be used to specify any number of contrasts. E.g. for levels: **1 2 3 4**

1) level 1 to level 3: 1 0 -1 0

2) level 2 to levels 1 and 4: -1/2 1 0 -1/2

3) levels 1 and 2 to levels 3 and 4: -1/2 -1/2 1/2 1/2

- The levels associated with the contrast coefficients with opposite signs are being compared.
  - The mean of the dependent variable is multiplied by the contrast coefficient
- Levels given a 0 are not involved in the comparison: they are multiplied by zero and "dropped out."

Specifying Multiple Contrasts

- Contrast coefficients must sum to zero. If they don't, the contrast is not estimable and you will get an error message.
- Which level of the categorical variable is assigned a positive or negative value is not terribly important
  - 1 0 -1 0 is equivalent to -1 0 1 0, but the sign of the regression coefficient would change
- In the 2nd and 3rd comparisons, the fractions sum to one (or minus one), but this is not necessary
  - While -1/2 1 0 -1/2 and -1 2 0 -1 both will give you the same t-value and p-value for the regression coefficient, the regression coefficients themselves would be different, as would their interpretation.

Implementing Multiple Contrasts

```
> mat = matrix(c(1, 0, -1, 0, -1/2, 1, 0, -1/2, -
  1/2, -1/2, 1/2, 1/2), ncol = 3)
> mat
     [,1] [,2] [,3]
[1,]    1 -0.5 -0.5
[2,]    0  1.0 -0.5
[3,]   -1  0.0  0.5
[4,]    0 -0.5  0.5
> my.contrasts = mat %*% solve(t(mat) %*% mat)
> my.contrasts
     [,1] [,2] [,3]
[1,] -0.5   -1 -1.5
[2,]  0.5    1  0.5
[3,] -1.5   -1 -1.5
[4,]  1.5    1  2.5
> contrasts(hsb2$race.f) = my.contrasts
```

Contrast Coding Wildcat Data

- R regression with the lmer function
  - Treatment Coding vs. Contrast Coding
- IV: native language - 3 levels
  - English (native)
  - Chinese (non-native)
  - Korean (non-native)
- Ordered the levels: English, Chinese, Korean

Treatment Coding

Input:

```
> langCompare.lmer =lmer(duration~lang+
   (1|Subject), data=myData)
```

Output:

	Estimate	Std. Error	t value
langChinese	0.025920	0.002384	10.872
langKorean	-0.004416	0.002091	-2.112

- Compares:
  - English vs. Chinese (langChinese)
  - English vs. Korean (langKorean)

Treatment Coding Matrix

- By default, R uses contr.treatment for unordered factors
- The first level of the factor is the baseline (here, English, so that the contrast matrix is all zeroes in the English row)

	Chinese	Korean
English	0	0
Chinese	1	0
Korean	0	1

Contrast Coding

Input:

> contrasts(myData$lang) = c(1, -.5, -.5)

- Compares the native (English) group to the non-native (Chinese and Korean) groups

> langCompare.lmer =lmer(duration~lang+ (1|Subject), data=myData)

Output:

	Estimate	Std. Error	t value
lang1	0.10002	0.010113	11.242
lang2	-0.00046	0.639887	1.388

Contrast Coding Matrix

- If too few [entries for the contrast matrix] are supplied, a suitable contrast matrix is created by extending value after ensuring its columns are contrasts (orthogonal to the constant term) and not collinear.

	[,1]	[,2]
English	1.0	-5.551115e-17
Chinese	-0.5	-7.071068e-01
Korean	-0.5	7.071068e-01

Interpreting Contrast Coding Outputs

	Estimate	Std. Error	t value
lang1	0.10002	0.010113	11.242
lang2	-0.00046	0.639887	1.388

- lang1: compares native (English) and non-native (Chinese and Korean) groups
- lang2: compares Chinese and Korean groups

Explanation of Contrast Coding 1

- Ignoring speaker-specific effects, the predicted mean for a given language is the intercept plus the dot product of the language's contrast-matrix representation with the coefficients for the language factor.
- Since the two models are equivalent, their predicted means are the same for each language.

Explanation of Contrast Coding 2

- The contrast matrix has columns summing to 0

  $\rightarrow$ The intercept can loosely be considered the predicted grand mean
- The coefficient for lang1 is the difference between
  - (a) the intercept and the English mean, and
  - (b) twice the difference between the intercept and the average of the Chinese and Korean means
- The coefficient for lang2 is the difference between Chinese and Korean divided by the square root of two
- These two coefficients operate on different scales, as reflected by the fact that the two columns of new.contrasts are vectors of different lengths

Further References

- Courtesy of Roger Levy
- Some useful information in:
  - Chambers & Hastie 1991, Section 2.3.2
  - Venables & Ripley 2002, Section 6.2
  - Healy 2000 ("Matrices for Statistics")