Categorical Variables in Regression Analyses

Maureen Gillespie

Northeastern University

May 3rd, 2010

References

Cohen, & Cohen, (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.*

- Especially chapters 8 & 9

Kaufman, D. & Sweet, R. (1974). Contrast coding in least squares regression analysis. *American Educational Research Journal*, *11*, 359–377.

Serlin, R. C., & Levin, J. R. (1985). Teaching how to derive directly interpretable coding schemes for multiple regression analysis. *Journal of Educational Statistics*, *10*, 223–238.

Wendorf, C. A. (2004). Primer on multiple regression coding: Common forms and the additional case of repeated contrasts. *Understanding Statistics*, *3*, 47–57

How do we treat categorical variables in regression?

- As sets of IVs (code variables)
  - Together they represent the full information from original categories.

How do we treat categorical variables in regression?

- As sets of IVs (code variables)
  - Together they represent the full information from original categories.
- Multiple ways to set up code variables
  - Different ways test different predictions
  - These are essentially planned comparisons

How many coding variables are necessary?

For any grouped/non-continuous IV (**G**) with some number of levels (g), g - 1 coding variables are needed to represent **G**.

- 4 levels $\rightarrow$ 3 coding variables (C_1, C_2, C_3)
- 3 levels $\rightarrow$ 2 coding variables (C_1, C_2)
- 2 levels $\rightarrow$ 1 coding variables (C_1)

NB:

g -1 = # of degrees of freedom (df) of **G**

Recap

- A categorical variable with g levels is represented by $g - 1$ coding variables, which means $g - 1$ coefficients to interpret.

Recap

- A categorical variable with g levels is represented by $g - 1$ coding variables, which means $g - 1$ coefficients to interpret.
  - The coefficients represent different comparisons under different coding schemes.
  - Overall model fit is the same regardless of coding scheme.

How do we represent the coding variables?

Common coding systems

- Treatment/Dummy Coding
- Effects/Sum Coding
- Planned/User-Defined/Contrast Coding (e.g.,Helmert)
- Polynomial Coding

NB:

The choice of your coding scheme affects the interpretation of the results for each **individual coding variable**; *however, it does not change the* **overall effect** *of the set of coding variables (i.e., model fit and related statistics will not be affected).*

Example Data Set

Lexical Decision Task

- Word status
  - /smok/ = word
  - /plok/ = phonologically legal nonword
  - /lbok/ = phonologically illegal nonword
- Task: Press button if the item sounds like an English word.
- DV: RT of response.

```
library(languageR)
################ Reads in Data ##############################
d<-read.table("fakedata.txt", header=TRUE)
#renames factors
d$IV1<-ifelse(d$IV1==1, "silent", "noise")
d$IV2<-ifelse(d$IV2==2, "word", ifelse(d$IV2==3, "legal",
"illegal"))

d$NoiseCond<-as.factor(d$IV1)
d$WordCond<-as.factor(d$IV2)
d$WordCond<-as.factor(d$WordCond)
d$NoiseCond<-as.factor(d$NoiseCond)
d$Freq<-as.numeric(d$Freq)
```

Check the structure of the data file by typing: head(d)

Example Data Set

Lexical Decision Task

- Does word status affect the time to make responses?

We'll run linear (and logistic) mixed-effect models testing this general question with different coding schemes.

- *One fixed effect (WordCond) and two random effects (Subject and Item intercepts)*

Treatment Coding

Compares other groups to a reference group.

- Considerations for choosing a reference group
  - Useful comparison (e.g., control, predicted highest or lowest)
  - Well-defined group (e.g., not a catch-all category)
  - Should **not** have small *n* compared to other groups
- Intercept represents the reference group mean.

Treatment Coding

Imagine the question we're interested in is whether responses to each of the nonword conditions differ from the *word* condition.

So, if this is our question, what level should we choose as a reference group?

Treatment Coding

Imagine the question we're interested in is whether responses to each of the nonword conditions differ from the *word* condition.

We should choose *word* as our reference group.

- Reference group receives a value of 0 for all coding variables (C_i)
- Each other level receives 1 in one of the coding variables

Levels	$\mathbf{C}_1$	$\mathbf{C}_2$
word	0	0
legal	1	0
illegal	0	1

$\mathbf{C}_1$ *tests legal against word*
$\mathbf{C}_2$ *tests illegal against word*

```
########### Sets up Treatment Coding ############################

# R automatically assigns levels alphabetically, this isn't always
what you'll want, so you can reassign the order of the levels as shown
below...

d$WordCond.Treatment<-d$WordCond

d$WordCond.Treatment<-factor(d$WordCond.Treatment,
levels=c("word","legal","illegal")) #reorders levels to put "word" in
baseline position (1st in list)

# R's default is to set coding scheme to Treatment, so here you don't
need to do anything else now that the levels are ordered
appropriately.

#More generally, if you just want to specify which level is the
baseline you can do the following:  contrasts(d$WordCond.Treatment)<-
contr.treatment(3, base=3)
#This says set the contrasts to treatment coding with 3 levels, with
the 3rd level being the base condition


lin.Treatment<-lmer(RT ~ WordCond.Treatment + (1|Subject) + (1|Item),
data=d) #linear model
```

Output from Treatment Coding Linear Regression Model

```
Fixed effects:
                           Estimate Std. Error t value
(Intercept)                  733.01      10.93   67.04
WordCond.Treatmentlegal      235.89      15.19   15.53
WordCond.Treatmentillegal    582.51      15.19   38.34
```

Output from Treatment Coding Linear Regression Model

```
Fixed effects:
                          Estimate Std. Error t value
(Intercept)                 733.01      10.93   67.04
WordCond.Treatmentlegal     235.89      15.19   15.53
WordCond.Treatmentillegal   582.51      15.19   38.34
```

- **Intercept**: English word mean RT is 733ms.

Output from Treatment Coding Linear Regression Model

```
Fixed effects:
                          Estimate Std. Error t value
(Intercept)                 733.01      10.93   67.04
WordCond.Treatmentlegal     235.89      15.19   15.53
WordCond.Treatmentillegal   582.51      15.19   38.34
```

- **Intercept**: English word mean RT is 733ms.
- $\mathbf{C}_1$: Legal nonwords are responded to 236ms slower than English words.

Output from Treatment Coding Linear Regression Model

```
Fixed effects:
                          Estimate Std. Error t value
(Intercept)                 733.01      10.93   67.04
WordCond.Treatmentlegal     235.89      15.19   15.53
WordCond.Treatmentillegal   582.51      15.19   38.34
```

- **Intercept**: English word mean RT is 733ms.
- $\mathbf{C}_1$: Legal nonwords are responded to 236ms slower than English words.
- $\mathbf{C}_2$: Illegal nonwords are responded to 583ms slower than English words.

Example 1

Now, let's imagine that you wanted to see RTs for phonologically legal items differ from the RTs for phonologically illegal items.

- Choose a base group that tests this question.
- Set up this coding scheme in R.
- Run the model and interpret the coefficients.

Output for Example 1

```
Fixed effects:
                            Estimate Std. Error t value
(Intercept)                  1315.52      10.93  120.31
WordCond.Treatment.Llegal    -346.62      15.19  -22.81
WordCond.Treatment.Lword     -582.51      15.19  -38.34
```

Output for Example 1

```
Fixed effects:
                            Estimate Std. Error t value
(Intercept)                  1315.52      10.93  120.31
WordCond.Treatment.Llegal    -346.62      15.19  -22.81
WordCond.Treatment.Lword     -582.51      15.19  -38.34
```

- **Intercept**: Illegal nonword mean RT is 1315ms.

Output for Example 1

```
Fixed effects:
                            Estimate Std. Error t value
(Intercept)                  1315.52      10.93  120.31
WordCond.Treatment.Llegal    -346.62      15.19  -22.81
WordCond.Treatment.Lword     -582.51      15.19  -38.34
```

- **Intercept**: Illegal nonword mean RT is 1315ms.
- $\mathbf{C}_1$: Legal nonwords are responded to 347ms faster than illegal nonwords.

Output for Example 1

```
Fixed effects:
                            Estimate Std. Error t value
(Intercept)                  1315.52      10.93  120.31
WordCond.Treatment.Llegal    -346.62      15.19  -22.81
WordCond.Treatment.Lword     -582.51      15.19  -38.34
```

- **Intercept**: Illegal nonword mean RT is 1315ms.
- $\mathbf{C}_1$: Legal nonwords are responded to 347ms faster than illegal nonwords.
- $\mathbf{C}_2$: Words are responded to 583ms faster than English words.

Effects Coding

Compares mean of a single group to the grand mean.

- Usually useful for unordered experimental groups
- *Base group* is chosen
  - Choose "least" interesting group
- Sum of the contrast weights of the coding variables always equals 0.
- Intercept represents the grand mean.

Effects Coding

Imagine that we choose *word* as our base group.

- Base group receives a value of -1 for all coding variables (C_i)
- Each other level receives 1 in one of the coding variables

Levels	$\mathbf{C}_1$	$\mathbf{C}_2$
word	-1	-1
legal	0	1
illegal	1	0

$\mathbf{C}_1$ *is the difference between illegal and grand mean.*
$\mathbf{C}_2$ *is the difference between the legal and grand mean.*

```
########### Sets up Effects Coding ##############################

d$WordCond.Effects<-d$WordCond

# If you want to make your outputs more readable (easier to figure
out what each Ci is doing), then you can manually create the
contrast matrix

contrasts(d$WordCond.Effects)<-cbind("illegal"= c(1, 0, -1),
"legal"= c(0, 1, -1))  #renames Cis to give indication of what is
being tested... C1 = illegal vs. grandmean, C2= legal vs.grandmean

#The basic command to create effects coding variables is
contr.sum(n), where n = the number of levels of your factor.

lin.Effects<-lmer(RT ~ WordCond.Effects + (1|Subject) + (1|Item),
data=d) #output is exactly the same as before, but now the levels
of output give description of contrasts
```

Output from Effects Coding Linear Regression Model

```
Fixed effects:
                          Estimate Std. Error t value
(Intercept)               1005.806      6.527  154.09
WordCond.Effectsillegal    309.712      8.772   35.31
WordCond.Effectslegal      -36.911      8.772   -4.21
```

Output from Effects Coding Linear Regression Model

```
Fixed effects:
                          Estimate Std. Error t value
(Intercept)               1005.806      6.527  154.09
WordCond.Effectsillegal    309.712      8.772   35.31
WordCond.Effectslegal      -36.911      8.772   -4.21
```

- **Intercept**: Grand mean RT is 1006ms.

Output from Effects Coding Linear Regression Model

```
Fixed effects:
                          Estimate Std. Error t value
(Intercept)               1005.806      6.527  154.09
WordCond.Effectsillegal    309.712      8.772   35.31
WordCond.Effectslegal      -36.911      8.772   -4.21
```

- **Intercept**: Grand mean RT is 1006ms.
- $\mathbf{C}_1$: Illegal nonwords are responded to $\sim$ 310ms slower than the grand mean.

Output from Effects Coding Linear Regression Model

```
Fixed effects:
                          Estimate Std. Error t value
(Intercept)               1005.806      6.527  154.09
WordCond.Effectsillegal    309.712      8.772   35.31
WordCond.Effectslegal      -36.911      8.772   -4.21
```

- **Intercept**: Grand mean RT is 1006ms.
- $\mathbf{C}_1$: Illegal nonwords are responded to $\sim$ 310ms slower than the grand mean.
- $\mathbf{C}_2$: Legal nonwords are responded to $\sim$ 37ms faster than the grand mean.

Orthogonal Contrast Coding

Goal of these coding systems is to allow each coding variable (C_i) to

- capture unique portions of the variance (i.e., orthogonal).
- test specific, theory-guided hypotheses (i.e., planned comparisons).

Constructing Orthogonal Contrast Codes (Cohen & Cohen, 1983)

- **Rule 1.** The sum of the weights across each code variable (C_i) must equal 0.

Constructing Orthogonal Contrast Codes (Cohen & Cohen, 1983)

- **Rule 1.** The sum of the weights across each code variable (C_i) must equal 0.
- **Rule 2.** The sum of the products of each pair of code variable (C_1, C_2) must equal 0.
  - When group sizes are equal, this ensures that contrast codes are orthogonal (i.e., do not capture overlapping portions of the variance).

Constructing Orthogonal Contrast Codes (Cohen & Cohen, 1983)

- **Rule 1.** The sum of the weights across each code variable (C_i) must equal 0.
- **Rule 2.** The sum of the products of each pair of code variable (C_1, C_2) must equal 0.
  - When group sizes are equal, this ensures that contrast codes are orthogonal (i.e., do not capture overlapping portions of the variance).
- **Rule/Suggestion 3.** The difference between the value of the set of positive weights and the value of the set of negatives weights should equal 1.
  - Allows each unstandardized β to correspond to the difference between the unweighted means of the groups involved in the contrast.

Ordered categorical variables

We often want to know whether levels of our independent variables are ordered.

- We could have a hypothesis that RT increases as "wordiness" decreases.

Helmert Coding

Tests one level of a factor against all previous levels.

- Useful for ordinal variables

Helmert Coding

Tests one level of a factor against all previous levels.

- Useful for ordinal variables
- Example comparisons
  - Does Level 1 differ from Level 2?
  - Does Level 1 differ from the mean of Levels 2 & 3?

Helmert Coding

Tests one level of a factor against all previous levels.

- Useful for ordinal variables
- Example comparisons
  - Does Level 1 differ from Level 2?
  - Does Level 1 differ from the mean of Levels 2 & 3?
- Intercept represents the grand mean.

Helmert Coding (Regression-style)

Are listeners sensitive to phonotactics of nonwords such that they more quickly perceive phonologically legal nonwords as words than phonologically illegal nonwords?

Are real English words more quickly perceived as words than nonwords?

Levels	$\mathbf{C}_1$	$\mathbf{C}_2$
word	0	2/3
legal	1/2	-1/3
illegal	-1/2	-1/3

$\mathbf{C}_1$ *tests legal against illegal*
$\mathbf{C}_2$ *tests word against mean of legal and illegal (i.e., word vs. nonword)*

NB:

R does not automatically assign weights that satisfy Rule/Suggestion 3.

```
########### Sets up Helmert Coding ##############################

##Regression style###


# Specifically, let's make each Ci equal the difference between
the means we're testing.
d$WordCond.Helm.Reg<-d$WordCond

contrasts(d$WordCond.Helm.Reg)<-cbind("leg.vs.ill"= c(-.5, .5,
0),"word.vs.nons"=c (-(1/3), -(1/3), (2/3)) )  #renames Cis to
give indication of what is being tested... C1 = illegal vs. legal,
C2= word vs.nonwords(mean of other two levels)


lin.Helm.Reg<-lmer(RT ~ WordCond.Helm.Reg + (1|Subject) + (1|
Item), data=d) # This model allows for directly interpretable Bs -
difference between conditions
```

```
Fixed effects:
                                Estimate Std. Error t value
(Intercept)                     1005.806      6.527  154.09
WordCond.Helm.Regleg.vs.ill     -346.622     15.193  -22.81
WordCond.Helm.Regword.vs.nons   -409.202     13.158  -31.10
```

$\mathbf{C}_1$: *Phonologically legal nonwords are responded to 346ms faster than phonologically illegal nonwords.*

$\mathbf{C}_2$: *English words are responded to 409ms faster than nonwords.*

Example 3

We can use any of these coding schemes for logistic models.

- Run a logistic regression using our accuracy measure as the dependent variable (d$Response).

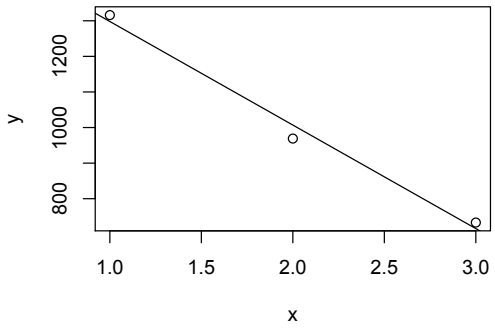
Output of Example 3

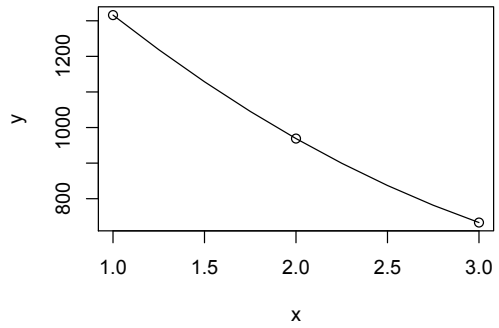
```
Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  0.02728    0.16085    0.17    0.865
WordCond.Helm.Regleg.vs.ill  2.02804    0.17549   11.56   <2e-16 ***
WordCond.Helm.Regword.vs.nons 1.83176   0.15267   12.00   <2e-16 ***
---
```

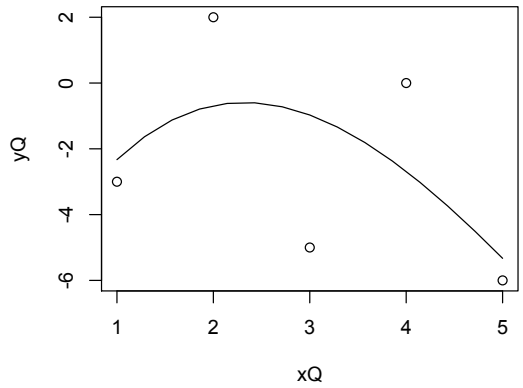
C$_1$: *Odds of a "word" response for phonologically legal nonwords are 7.6 ($e^{2.03}$) times higher than the odds for phonologically illegal nonwords.*
C$_2$: *Odds of a "word" response for English words are 6.2 ($e^{1.83}$) times higher than the odds for nonwords.*

Polynomial Coding

What if we care about the shape of the effect over a range of ordered levels of our independent variable, rather than differences between group means?







Polynomial Coding

How do we model trends in ordered categorical variables?

- Linear trend?
- Quadratic trend?
- Higher-level trends?

Can test for g - 1 higher-order trends.

- 2-level factor: Linear (X^1)
- 3-level factor: Linear, Quadratic (X^2)
- 4-level factor: Linear, Quadratic, Cubic (X^3)

NB:

Orthogonal polynomial contrasts can be automatically generated by R for any number of levels using the function contr.poly(n), where n = number of levels of your factor.

Polynomial Coding

How do we model trends in ordered categorical variables?

```
                  .L          .Q
illegal -7.071068e-01  0.4082483
legal   -7.850462e-17 -0.8164966
word     7.071068e-01  0.4082483
```

C_1 (.L) *tests if there is a linear component.*
C_2 (.Q) *tests if there is a quadratic component.*

```
d$WordCond.Poly<-d$WordCond

contrasts(d$WordCond.Poly)<-contr.poly(3) #This specifies that you
want to do polynomial coding (tests linear and quadratic
components for this example with 3-levels)

#Orthogonal coding scheme, but interpretations of Cis are
different because you're not testing for differences among group
means

lin.Poly<-lmer(RT ~ WordCond.Poly + (1|Subject) + (1|Item),
data=d)


# Output shows .L for linear component and .Q for quadratic.  If
the .L coeff is significant, then the regression requires a linear
component, if the .Q coeff is significant, the effect has a
quadratic component.
```

But what about the main effect?

If you want to get an estimate of the main effect of a multi-level categorical variable, you can use the function aovlmer.fnc().

```
Analysis of Variance Table
                Df   Sum Sq  Mean Sq F value      F      Df2         p
WordCond.Poly    2 65934450 32967225  743.82 743.82     1149         0

Analysis of Variance Table
                  Df   Sum Sq  Mean Sq F value      F      Df2         p
WordCond.Treatment  2 65934450 32967225  743.82 743.82     1149         0
```

Overall effect of the variable is the same, regardless of coding scheme, but individual coding variables (C_i) will differ

- Treatment $C_1 \neq$ Effects $C_1 \neq$ Helmert $C_1 \neq$ Polynomial C_1

Each **complete set** *of coding variables captures the* **same** *overall proportion of the variance in the DV, but the interpretation of each* **individual coding variable** *is different under different coding schemes.*

General Overview

The choice of your coding scheme affects the interpretation of the results for each **individual coding variable**; however, it does not change the **overall effect** of the set of coding variables (i.e., model fit, and related statistics, will not be affected).