Categorical Predictors 1

Paul E. Johnson[1] [2]

[1]Department of Political Science

[2]Center for Research Methods and Data Analysis, University of Kansas

2014

Introduction

- Basics: Before I get too carried away
- Categorical Coding: Which Dummy is Right for you?
- Differences among approaches are Superficial

Outline

Let's Talk About Sex

- Sex is coded "M" for male or "F" for female
- "manually" create two dummy variables, "femd" and "maled"
- These are numeric, 0 or 1 (or maybe -1 and 1).
- In SAS (or Stata), one then fits a model using "femd" or "maled" as a predictor.

id	constant	sex	femd	maled
1	1	M	0	1
2	1	F	1	0
3	1	F	1	0
4	1	M	0	1
⋮			⋮	

What will R do if...

- ```
  lm (y ~ sex)
  ```

  fits

  - (implicitly) asks for an intercept, plus
  - an "intercept shift" parameter for a contrast variable for males it calls "sexM".
  - R automatically creates a "contrast" variable, a 0, 1 "dummy" variable for male
  ```

Example: statusquo support in the 1988 Chile Data

```
library(car)
mod1 <- lm(statusquo ~ sex, data=Chile)
summary(mod1)
```

	M1	
	Estimate	(S.E.)
(Intercept)	0.066*	(0.027)
sexM	-0.134***	(0.039)
N	2683	
RMSE	0.998	
R^2	0.004	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

Sex Contrast Default and Interpretation

- R's design matrix that looks like this:

$$X = \begin{array}{cc} constant & sexM \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ & \vdots \end{array} \qquad (1)$$

- Why "M"? Female becomes "baseline" (in the intercept) because it is alphabetically first (can customize that)
- Same effect as user-created "maled" variable.
- fitted intercept represents the effect of "being human" (or "being in the data set")
- $\hat{b}_1 sexM$; the "difference" effect that distinguishes males from other humans
- Model's predicted value is $\widehat{statusquo_i} = \hat{b}_0 + \hat{b}_1 sexM$, so for Females predict $\hat{b}_0$ and for males predict $\hat{b}_0 + \hat{b}_1$.

Regression Equivalent to a "t-test for means"

The "t test for means" calculates the averages within groups and calculates a t value for the difference.

```
by(Chile$statusquo, Chile$sex, mean, na.rm = TRUE)
```

```
Chile$sex: F
[1] 0.06570627
------------------------------------------------------------

Chile$sex: M
[1] -0.06835453
```

```
t.test(statusquo ~ sex, var.equal=TRUE, data=Chile
   )
```

Regression Equivalent to a "t-test for means" ...

```
  Two Sample t-test

data:  statusquo by sex
t = 3.4779, df = 2681, p-value = 0.0005135
alternative hypothesis: true difference in means is
     not equal to 0
95 percent confidence interval:
 0.05847624 0.20964537
sample estimates:
mean in group F mean in group M
     0.06570627      -0.06835453
```

Note the Regression intercept and slope re-produce means as predicted values.

Outline

Occupation in the wages data set

- As provided, wages has occupation coded as a numeric variable.
-
1	2	3	4	5	6
Management	Sales	Clerical	Service	Professional	Other

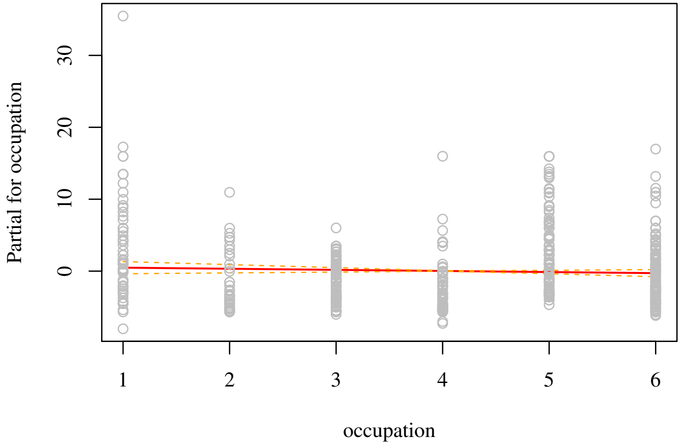
See Why it is Wrong to treat that as Numeric, Right?

```
mod1 <- lm(wage ~ occupation, data=dat)
```

	M1	
	Estimate	(S.E.)
(Intercept)	9.656***	(0.600)
occupation	-0.152	(0.134)
N	534	
RMSE	5.138	
R^2	0.002	

$*p \le 0.05 ** \ p \le 0.01 *** p \le 0.001$

Interpret that Termplot



Recode, Treat Occupation as A Categorical Variable

- Create a new “factor” variable occupationf, that assigns labels to the categories.
- When there are 6 occupational categories, the usual approach creates 5 “dummy variables”
- In R, those 5 dummy variables are created automatically, called “treatment contrasts”
- “first” level of factor (or designated level) is excluded, and rest of levels are “dummied up”

What is R Doing with "occupationf"?

- R's system of "factor" variables is intended to make this "automatic". Regression procedures create "contrasts" "on the fly".
- The factor "occupationf" is converted thus

	Sales	Clerical	Service	Professional	Other
Management	0	0	0	0	0
Sales	1	0	0	0	0
Clerical	0	1	0	0	0
Service	0	0	1	0	0
Professional	0	0	0	1	0
Other	0	0	0	0	1

- So the fitted model for 6 categories is

$$\widehat{wages}_i = \hat{b}_0+\hat{b}_1 Sales_i+\hat{b}_2 Clerical_i+\hat{b}_3 Service_i+\hat{b}_4 Professional_i+\hat{b}_5 Other_i \tag{2}$$

- Maybe I should make this easier to remember

$$\begin{aligned}\widehat{wages}_i &= \hat{b}_0 + \hat{b}_{Sales} Sales_i + \hat{b}_{Clerical} Clerical_i \\ &+\hat{b}_{Service} Service_i + \hat{b}_{Prof} Professional_i + \hat{b}_{Other} Other_i\end{aligned}$$

Fitted Regression Model with Categorical Predictor

	M1	
	Estimate	(S.E.)
(Intercept)	12.704***	(0.630)
occupationfSales	-5.111***	(0.986)
occupationfClerical	-5.281***	(0.789)
occupationfService	-6.167***	(0.813)
occupationfProfessional	-0.757	(0.778)
occupationfOther	-4.278***	(0.733)
N	534	
RMSE	4.675	
R^2	0.180	
adj R^2	0.173	

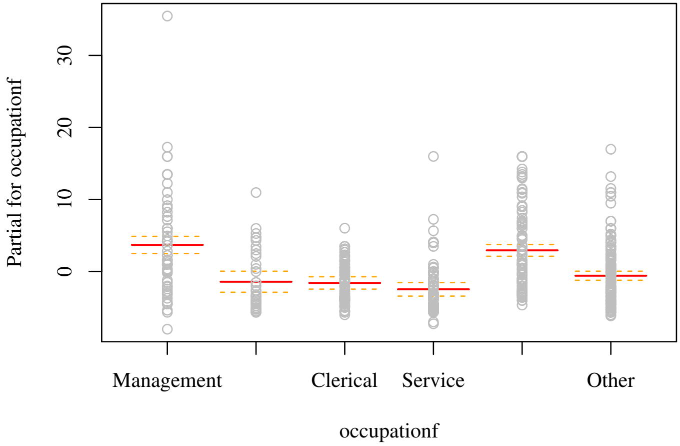
$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

Management is the "baseline". Calculate Predicted Values:

$\hat{y}_{Management} = \hat{b}_0 = 12.704 \qquad \hat{y}_{Sales} = \hat{b}_0 + \hat{b}_{Sales} = 12.704 - 5.11 = 7.59$

$\hat{y}_{Service} = 12.704 - 6.167 = 6.537$

Interpret that Termplot



Contrasts:

- The default treats the "lowest" score–the first "level"–as a "baseline" category.
  - Meaning: There is no "dummy" variable for that. It is "in" the intercept.
- All other categories are compared against that one.

Does the occupationf ”Belong” in the Model

- Obviously Yes: “occupationf” makes a difference–some categories matter
- Formally test with F test, where null is that none of the differences are non-zero.

$$H_0 : \hat{b}_{Sales} = \hat{b}_{Clerical} = \hat{b}_{Service} = \hat{b}_{Professional} = \hat{b}_{Other} = 0 \qquad (3)$$

- Compare the fitted model against a model that has only the intercept
- That’s the F test that is reported with most regression models.

```
summary(mod2)
```

Does the occupationf "Belong" in the Model ...

```
Call:
lm(formula = wage ~ occupationf, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-11.704  -3.041  -1.037   2.296  31.796

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                12.7040     0.6304  20.154  < 2e-16 ***
occupationfSales           -5.1114     0.9861  -5.183 3.11e-07 ***
occupationfClerical        -5.2814     0.7891  -6.693 5.59e-11 ***
occupationfService         -6.1665     0.8128  -7.587 1.49e-13 ***
occupationfProfessional    -0.7566     0.7781  -0.972    0.331
occupationfOther           -4.2775     0.7331  -5.835 9.40e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.675 on 528 degrees of freedom
Multiple R^2:  0.1803,  Adjusted R^2:  0.1725
F-statistic: 23.22 on 5 and 528 DF,  p-value: < 2.2e-16
```

Does the occupationf "Belong" in the Model

- R's anova function provides a conventional "analysis of variance table".

```
anova(mod2, test="F")
```

```
Analysis of Variance Table

Response: wage
             Df  Sum Sq Mean Sq F value    Pr(>F)
occupationf   5  2537.7  507.54  23.224 < 2.2e-16 ***
Residuals   528 11539.0   21.85
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Outline

1. Basics
   - Dichotomy
   - Multichotomy (Polychotomy?)
   - Simplify the Coding

2. Coding Schemes
   - G-1 is Over-rated
   - You Want G Parameters? You Got It!
   - Same True With G Categories

3. Effects Coding

But Do We Really Need All Those Parameters?

- Glance at the estimated slope coefficients.
- I suspect the middle 3 categories have “about the same” effect

Hypothesis Testing Procedure

- F test
- $H_0 : b_{sales} = b_{service} = b_{clerical}$
- Estimate "full" or "unrestricted" model with all of the category dummies included
- Estimate "partial" or "restricted" model with restriction imposed.
- Compare the fit, F test indicates whether estimates $\hat{b}_{sales}$, $\hat{b}_{service}$, $\hat{b}_{clerical}$, are "statistically significantly different" from one another.
- Slang: is "predictive power" lost by restriction?

Test $\hat{b}_{Sales} = \hat{b}_{Clerical} = \hat{b}_{Service}$

- Testing the restriction that the wage effect for three groups is achieved by recoding occupationf variable
- All "Sales" "Clerical" and "Service" observations re-coded 1 on new category "sales/clerical/service"

	M1	
	Estimate	(S.E.)
(Intercept)	12.704***	(0.630)
occupationf2sales/clerk/serv	-5.589***	(0.705)
occupationf2Professional	-0.757	(0.778)
occupationf2Other	-4.278***	(0.733)
N	534	
RMSE	4.675	
R^2	0.177	
adj R^2	0.172	

$*p \leq 0.05 ** p \leq 0.01 *** p \leq 0.001$

And the F test result is (drumroll please)

```
anova(mod3, mod2, test="F")
```

```
Analysis of Variance Table

Model 1: wage ~ occupationf2
Model 2: wage ~ occupationf
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    530 11584
2    528 11539  2    45.529 1.0417 0.3536
```

What if I merge "Management" and "Professional"?

- Appears to me $\hat{y}_{Professional}$ and $\hat{y}_{Management}$ are not all that different.
- Suppose $H_o : b_{Professional} = 0$ and $b_{sales} = b_{service} = b_{clerical}$
- Then we create an even simpler variable, which leads to 2 "dummy" variables

	sales/clerk/serv	Other
manag/prof	0	0
sales/clerk/serv	1	0
Other	0	1

And the Regression on that Simpler Set of Contrasts is

	M1	
	Estimate	(S.E.)
(Intercept)	12.207***	(0.370)
occupationf2sales/clerk/serv	-5.092***	(0.487)
occupationf2Other	-3.781***	(0.526)
N	534	
RMSE	4.675	
R^2	0.176	
adj R^2	0.172	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

And The F Test says

- Compare the "full" fitted model with all 5 category differences estimated
- With the restricted model

```
anova( mod4, mod2,  test="F")
```

```
Analysis of Variance Table

Model 1: wage ~ occupationf2
Model 2: wage ~ occupationf
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    531 11605
2    528 11539  3     66.19 1.0096 0.3881
```

Conclusion: Does not appear the model with 3 categories (intercept + 2 group contrasts) has a worse statistical fit.

Outline

What To Do with a G-Category Nominal Variable?

- If there are G categories,
- Texts usually say "regression can provide parameter estimates for G-1 categories"
- Strinctly Speaking, that's wrong.
    - It is only true if you include an Intercept in your regression
    - Drop the intercept, you can have G category estimates!

Lets Talk About Sex (again!)

- Recall, the data has a categorical "sex" (M or F) and we can create "dummy" variables for females and males.

id	constant	sex	femd	maled
1	1	M	0	1
2	1	F	1	0
3	1	F	1	0
4	1	M	0	1
⋮			⋮	

- You agree, don't you, that:
  - We get essentially the same model if we fit a dummy variable for "female" or for "male", right?
  - $\hat{y}_i = \hat{b}_0 + \hat{b}_1 \cdot femd_i$ treats "male" as baseline and $\hat{b}_1$ is the difference for females
  - $\hat{y}_i = \hat{b}_0 + \hat{b}_1 \cdot maled_i$ treats "female" as baseline and $\hat{b}_1$is the difference for males

Outline

Drop the Intercept? Intriguing!

- Drop the intercept? G categories -> G parameter estimates
- lm(y ~ -1 + sex) : fits no intercept, estimates parameters for both males and females

$$\begin{array}{cc} sexF & sexM \\ 0 & 1 \\ 1 & 0 \end{array} \tag{4}$$

- And that is “essentially the same model” as either of the others.

Problem comes back to Multicollinearity

- See why you can't estimate this:

```
lm(y~femd+maled)
```

- R automatically inserts an "intercept" coefficient for you, so this is really

```
lm(y~1+femd+maled)
```

- Leading to the design matrix on right: perfect collinearity between constant, femd and maled
  - Your options:
    - include a constant and either femd or maled
    - remove the constant and estimate femd and maled

constant	femd	maled
1	0	1
1	1	0
1	1	0
1	0	1

Better Check that with the Chile Data

- Traditional model, sexM

```
chile1M <- lm(statusquo ~ sex, data=Chile)
```

- Traditional model, sexF

```
Chile$sex <- relevel(Chile$sex, ref="M")
chile1F <- lm(statusquo ~ sex, data=Chile)
```

- No Intercept Model

```
chile1NI <- lm(statusquo ~ -1 + sex, data=Chile)
```

3 Fits Side By Side

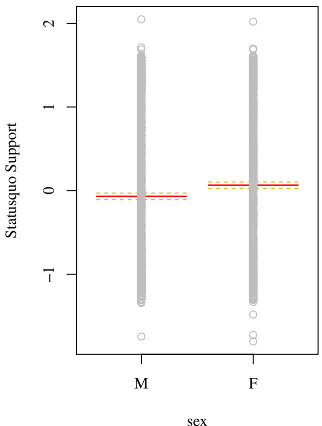
	M Estimate (S.E.)	F Estimate (S.E.)	No Int. Estimate (S.E.)
(Intercept)	0.066*	-0.068*	.
	(0.027)	(0.028)	
sexM	-0.134***	.	-0.068*
	(0.039)		(0.028)
sexF	.	0.134***	0.066*
		(0.039)	(0.027)
N	2683	2683	2683
RMSE	0.998	0.998	0.998
R^2	0.004	0.004	0.004
adj R^2	0.004	0.004	0.004

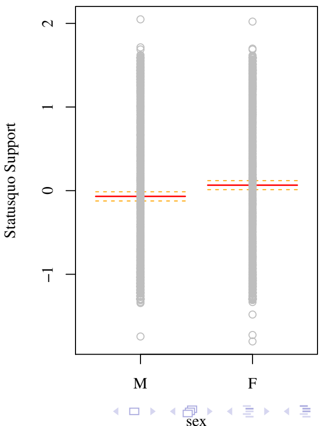
$*p \leq 0.05{**}\ p \leq 0.01{***}p \leq 0.001$

Vital: The Predicted Values Are IDENTICAL!

```
chile1F <- lm(statusquo ~ sex,
    data=Chile)
```

```
chile1NI <- lm(statusquo ~ -1
    + sex, data=Chile)
```





I mean Predictions are Completely IDENTICAL! Check the first few cases

```
head(predict(chile1F))
```

```
          1           2           3           4
                      5           6
-0.06835453 -0.06835453  0.06570627  0.06570627  0
  .06570627  0.06570627
```

```
head(predict(chile1NI))
```

```
          1           2           3           4
                      5           6
-0.06835453 -0.06835453  0.06570627  0.06570627  0
  .06570627  0.06570627
```

Outline

So, if a Categorical IV has 5 "levels" (as R would call them)

- We can estimate 4 parameters for levels and 1 for intercept
- Or we can suppress intercept and estimate 5 parameters for 5 levels

Treatment Contrasts==”dummy” codes

- Colloquial: Dummy Variable Coding
- R calls this “treatment contrasts”

id	Religion	Rel.Cath	Rel.Prot	Rel.Musl	Rel.Hindu	Rel.Other
1	Cath	1	0	0	0	0
2	Prot	0	1	0	0	0
3	Musl	0	0	1	0	0
4	Hindu	0	0	0	1	0
5	Other	0	0	0	0	1
6	⋮					

Regression with Treatment Contrasts

- $\hat{y}_i \sim \hat{b}_0 + \hat{b}_1 Rel.Prot_i + \hat{b}_2 Rel.Musl_i + \hat{b}_3 Rel.Hindu_i + \hat{b}_4 Rel.Other_i$
- "Catholic" is "left out?" Not really
- Predicted value for members of
    - Catholic is $\hat{b}_0$
    - Protestant is $\hat{b}_0 + \hat{b}_1$
    - Muslim is $\hat{b}_0 + \hat{b}_2$
    - Hindu is $\hat{b}_0 + \hat{b}_3$
    - Other is $\hat{b}_0 + \hat{b}_4$
- Interpret individual coefficients
    - $\hat{b}_1$: difference in predicted value for Protestant (as opposed to Catholic).
    - $\hat{b}_2$: difference in predicted value for Muslim (as compared against Catholic)

Any Group Can Serve as the Baseline

- Can make "Hindu" the baseline group.
- All estimates treat Hindu as "baseline" and other estimates are differences in prediction against Hindu category
- Model predictions and fit indices are still IDENTICAL to other "Catholic baseline" model.
- If there are no other predictors in the model, the $\hat{b}_j's$ are simply related to the observed group means (since predicted value is "mean" of y for category members).

Remember $\hat{y}$ is the same, no matter how you code these Predictor Contrasts

- Changing "dummy codes" or "baseline group" alters the $\hat{b}$ estimates
- It does not alter the essential meaning of the model
- Like saying "I am average in height" and "my height is the average plus 0" or "my height is 36 inches plus one-half of the average"

Effects Coding (Unweighted)

- Terminology is "new to me" in Cohen, et al.
- Re-code the religion variable like so (for "omitted" category, put -1 all the way across)

id	Religion	Rel.Cath	Rel.Prot	Rel.Musl	Rel.Hindu	Rel.Other
1	Cath	-1	-1	-1	-1	-1
2	Prot	0	1	0	0	0
3	Musl	0	0	1	0	0
4	Hindu	0	0	0	1	0
5	Other	0	0	0	0	1
6	⋮					

- Called "sum-to-zero" contrasts in other contexts.
- We will fit a regression that does not include *Rel.Cath*
  $\hat{y}_i \sim \hat{b}_0 + \hat{b}_1 Rel.Prot_i + \hat{b}_2 Rel.Musl_i + \hat{b}_3 Rel.Hindu_i + \hat{b}_4 Rel.Other_i$
- Still get $\hat{b}$'s as comparisons, but now comparing against a different baseline.

Design Matrix

The "design matrix":

Const	Cath	P	M	H	Oth
1	-1	-1	-1	-1	-1
1	0	1	0	0	0
1	0	0	1	0	0
1	0	0	0	1	0
1	0	0	0	0	1
⋮					

But "Cath" is omitted from the fitted report

- Every "row" gets
  - a 1 for its "own" group
  - Except Catholics, who get -1
- The -1 basically "pushes" the estimated intercept
- The other coefficients adjust accordingly to produce same predicted values.

Where does the Intercept get pushed to?

- Answer: Intercept=mean of group means on y

$$\hat{b}_0 = \frac{1}{5}\{\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3 + \bar{Y}_4 + \bar{Y}_5\} \tag{5}$$

- Called "unweighted effects coding" because the means of the groups are averaged, no matter how many observations there are in each group.
- In order to believe that, I had to run some examples.

Chile Regions: First get the means

- The mean values of "statusquo" for the regions are

```
  region           x
1      C -0.02983546
2      M  0.28677120
3      N  0.13556488
4      S  0.16496487
5     SA -0.17955745
```

- Now calculate the "mean of the means" (no weights)

```
[1] 0.07558161
```

0.076 is a "magic number". Watch out for it later

Suppress the Intercept: Estimate 5 Params for 5 Regions

```
modr1 <- lm( statusquo ~ -1 + region , data=Chile)
outreg(modr1 , tight=FALSE, showAIC=F)
```

	M1	
	Estimate	(S.E.)
regionC	-0.030	(0.040)
regionM	0.287**	(0.099)
regionN	0.136*	(0.055)
regionS	0.165***	(0.037)
regionSA	-0.180***	(0.032)
N	2683	
RMSE	0.989	
R^2	0.024	
adj R^2	0.022	

$*p \leq 0.05 ** \; p \leq 0.01 *** p \leq 0.001$

Include the Intercept, Estimate (default) Treatment Contrasts

```
modr2 <- lm( statusquo ~ region , data=Chile , x=T,
    y=T)
outreg(modr2 , tight=FALSE, showAIC=F)
```

	M1	
	Estimate	(S.E.)
(Intercept)	-0.030	(0.040)
regionM	0.317**	(0.107)
regionN	0.165*	(0.068)
regionS	0.195***	(0.055)
regionSA	-0.150**	(0.052)
N	2683	
RMSE	0.989	
R^2	0.024	
adj R^2	0.023	

$*p \le 0.05 ** \ p \le 0.01 *** p \le 0.001$

Those Default Contrasts Were

```
contrasts(Chile$region)
```

```
   M N S SA
C  0 0 0  0
M  1 0 0  0
N  0 1 0  0
S  0 0 1  0
SA 0 0 0  1
```

Ask R to use "sum-to-zero" contrasts (aka Unweighted Effects)

```
options(contrasts=c("contr.sum", "contr.poly"))
contrasts(Chile$region)
```

```
   [,1] [,2] [,3] [,4]
C     1    0    0    0
M     0    1    0    0
N     0    0    1    0
S     0    0    0    1
SA   -1   -1   -1   -1
```

- Note, the default makes the "last" category, SA, the reference category. Will have to fix that later.

Fitted model with Effects Contrasts

	M1	
	Estimate	(S.E.)
(Intercept)	0.076**	(0.026)
region1	-0.105**	(0.041)
region2	0.211**	(0.081)
region3	0.060	(0.050)
region4	0.089*	(0.039)
N	2683	
RMSE	0.989	
R^2	0.024	
adj R^2	0.023	

$*p \leq 0.05 ** \; p \leq 0.01 *** p \leq 0.001$

- Unfortunately, we lose the region labels here, but they are 1=C, 2=M, 3=N, 4=S

I Had Trouble figuring this Out

- Some patience required :)
- Note the Effects Coding intercept is 0.076, same as "mean of category means"

- Calculate the difference between the observed means and 0.076

```
  region           x        diff
1      C −0.02983546 −0.10541707
2      M  0.28677120  0.21118959
3      N  0.13556488  0.05998327
4      S  0.16496487  0.08938326
5     SA −0.17955745 −0.25513905
```

Note those differences exactly reproduce the $\hat{b}$ estimates from the unweighted effects model.

I wish C were the Omitted Category

- Create a new factor "region2" in which levels are ordered (M, N, S, SA, C)
- That forces values for cases in C to -1 for all contrasts

```
   [,1] [,2] [,3] [,4]
M     1    0    0    0
N     0    1    0    0
S     0    0    1    0
SA    0    0    0    1
C    -1   -1   -1   -1
```

Re-fit with "C" as the reference

	M1	
	Estimate	(S.E.)
(Intercept)	0.076**	(0.026)
region21	0.211**	(0.081)
region22	0.060	(0.050)
region23	0.089*	(0.039)
region24	-0.255***	(0.036)
N	2683	
RMSE	0.989	
R^2	0.024	
adj R^2	0.023	

$*p \leq 0.05{**}\ p \leq 0.01{*}{*}{*}p \leq 0.001$

Interpretation benefit to the $\hat{b}$'s

- One can scan down the parameter estimates to see if one category is above the unweighted mean
- Unclear to me why one would want to do that, but one can, if one wants to

But they are all Fundamentally the same

No Intercept

	M1 Estimate (S.E.)
regionC	-0.030 (0.040)
regionM	0.287** (0.099)
regionN	0.136* (0.055)
regionS	0.165*** (0.037)
regionSA	-0.180*** (0.032)
N	2683
RMSE	0.989
R^2	0.024
adj R^2	0.022

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

Treatment

	M1 Estimate (S.E.)
(Intercept)	-0.030 (0.040)
regionM	0.317** (0.107)
regionN	0.165* (0.068)
regionS	0.195*** (0.055)
regionSA	-0.150** (0.052)
N	2683
RMSE	0.989
R^2	0.024
adj R^2	0.023

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

Effects

	M1 Estimate (S.E.)
(Intercept)	0.076** (0.026)
region21	0.211** (0.081)
region22	0.060 (0.050)
region23	0.089* (0.039)
region24	-0.255*** (0.036)
N	2683
RMSE	0.989
R^2	0.024
adj R^2	0.023

$*p \leq 0.05 ** \ p \leq 0.01 *** p$

Predicted Values for all Rows are Identical. Same, Equivalent, Interchangeable

- Note predicted values for all regions are same

```
  region       NoInt   Treatment     Effects
1      C -0.02983546 -0.02983546 -0.02983546
2      M  0.28677120  0.28677120  0.28677120
3      N  0.13556488  0.13556488  0.13556488
4      S  0.16496487  0.16496487  0.16496487
5     SA -0.17955745 -0.17955745 -0.17955745
```

- R's "all.equal" verifies that the predictions for each row in data are same.

```
all.equal(predict(modr1), predict(modr2), predict(
    modr3))
```

```
[1] TRUE
```

The Standard Errors of the $\hat{b}$ Only Appear to Differ

- The standard errors are different, but
- That's only because they are estimating different things!
- $Std.Err.(\hat{b})$ varies because each model reports an estimate of a different value
- The No Intercept model estimates a "total effect" value for each region
- The Treatment Contrast model estimates
  - one "total effect" for baseline
  - difference for each region against baseline
- Effects Contrasts estimate
  - one unweighted mean
  - differences for each region against that

Consider Region S

- No Intercept model $\hat{b}_S = 0.165$, $Std.Err(\hat{b}_S) = 0.037$
- Treatment Contrasts, $\hat{b}_S = 0.195$, $Std.Err(\hat{b}_s) = 0.055$
- Effects Contrasts, $\hat{b}_S = 0.089$, $Std.Err.(\hat{b}_S) = 0.039$
- From Treatment, can re-construct estimate for "total S region effect"

$$\hat{b}_0 + \hat{b}_S \; with \; Std.Err.(\sqrt{Var(\hat{b}_0) + Var(\hat{b}_S) + 2Cov(\hat{b}_0, \hat{b}_S)}) \quad (6)$$

  - Inserting values from the Covariance of the $\hat{b}$ from Treatment gives 0.037
- Do same with Effects Contrasts, get standard error of 0.037

My "Take Away" Message

- Regression is a "vehicle" with which to calculate predicted values
- Many equivalent "design matrices" can be used to calculate same predicted values
- Comfort with one method or its estimates b's drives the selection of one's approach. There is no "real" methodological difference between the two.
- Often choose approach so that "free t-tests" with regression output are testing the most meaningful questions.