CATEGORICAL VARIABLES IN LINEAR REGRESSION MODELS

Recall NC Mercury fish data.

- STATION is a categorical variable
- $0, 1, \ldots, 15$ labels
- Default is to treat integer data as numerical
- use `factor(STATION)` in model formula
  `lm(MERCURY ~ factor(STATION), data=fish)`
- Can coerce to be a `factor` with `as.factor()`
  `fish$STATION <- as.factor(fish$STATION)`
  `lm(MERCURY ~ STATION, data=fish)`
  `aov(MERCURY ~ STATION, data=fish)`

FACTOR CONTRASTS

A `factor` variable in a model formula is expanded to create a design matrix using *contrasts*

- natural contrasts are treatment contrasts `contr.treatment()`
  - series of K-1 indicator or DUMMY variables for each factor level
  - excludes the first level (effect is estimated by the intercept)
- Sum to Zero contrasts `contr.sum()`
- Helmert Contrasts `contr.helmert()`
- for ordered factors orthogonal polynomials `contr.poly`

DEFAULT CONTRASTS

Use `options()$contrasts` to view and/or replace default contrasts in R/S

- in R

```
> options()$contrasts
        unordered           ordered
"contr.treatment"      "contr.poly"
```

- In S-Plus

```
> options()$contrasts
          factor      ordered
 "contr.helmert" "contr.poly"
> options(contrasts=c("contr.treatment", "contr.poly"))
# replaces default with contr.treatment
```

CHANGING CONTRASTS

Rather than changing the global default, for a factor variable, the default contrasts can be replaced by changing the *attributes* of the factor variable

```
> contrasts(fish$STATION) <- "contr.helmert"
> attributes(fish$STATION)
$levels
 [1] "0"  "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"
  "10" "11" "12" "13" "14" "15"

$class
[1] "factor"

$contrasts
[1] "contr.helmert"
```

COEFFICIENTS

- different contrasts lead to different design matrices for representing the factor
- different coefficient estimates and different standard errors
- Careful with interpretation of coefficient estimates!
- sum of squares and ANOVA decompositions for the factor are the same
- testing for factor inclusion is the same

```
> contrasts(fish$STATION) <- "contr.helmert"
> anova(lm(MERCURY ~ STATION, fish))
Analysis of Variance Table
Response: MERCURY
           Df Sum Sq Mean Sq F value    Pr(>F)
STATION    15 37.600   2.507  6.3671 2.100e-10 ***
Residuals 155 61.022   0.394

> contrasts(fish$STATION) <- "contr.treatment"
> anova(lm(MERCURY ~ STATION, fish))
Analysis of Variance Table
Response: MERCURY
           Df Sum Sq Mean Sq F value    Pr(>F)
STATION    15 37.600   2.507  6.3671 2.100e-10 ***
Residuals 155 61.022   0.394
```

```
> contrasts(fish$STATION) <- "contr.helmert"
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.218509   0.057402  21.228  < 2e-16 ***
STATION1    -0.074395   0.122566  -0.607 0.544754
STATION2     0.074055   0.070951   1.044 0.298229
STATION3     0.077085   0.055571   1.387 0.167391
STATION4    -0.009520   0.038068  -0.250 0.802850
STATION5    -0.051704   0.061812  -0.836 0.404175
STATION6     0.153426   0.046610   3.292 0.001233 **
STATION7    -0.058993   0.027179  -2.171 0.031489 *
   .            .          .         .       .
STATION12   -0.050464   0.014348  -3.517 0.000572 ***
STATION13    0.009108   0.022862   0.398 0.690891
STATION14    0.057842   0.012338   4.688 6.00e-06 ***
STATION15    0.059069   0.012403   4.763 4.36e-06 ***
```

```
> contrasts(fish$STATION) <- "contr.treatment"
> summary(lm(MERCURY ~ STATION, fish))
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.99300    0.19842   5.005 1.50e-06 ***
STATION1    -0.14879    0.24513  -0.607  0.54475
STATION2     0.14777    0.26392   0.560  0.57635
STATION3     0.30800    0.28060   1.098  0.27407
STATION4     0.02914    0.25979   0.112  0.91083
STATION5    -0.24300    0.41304  -0.588  0.55717
STATION6     1.08950    0.37120   2.935  0.00384 **
STATION7    -0.30300    0.28060  -1.080  0.28190
   .
STATION12   -0.50608    0.26392  -1.918  0.05701 .
STATION13    0.22700    0.37120   0.612  0.54175
STATION14    0.97623    0.26392   3.699  0.00030 ***
STATION15    1.11155    0.27415   4.054 7.93e-05 ***
```

```
> contrasts(fish$STATION) <- "contr.sum"
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.218509   0.057402  21.228  < 2e-16 ***
STATION1    -0.225509   0.194275  -1.161  0.24752
STATION2    -0.374299   0.146374  -2.557  0.01151 *
STATION3    -0.077740   0.172608  -0.450  0.65306
STATION4     0.082491   0.194275   0.425  0.67171
STATION5    -0.196366   0.167035  -1.176  0.24156
STATION6    -0.468509   0.343688  -1.363  0.17480
STATION7     0.863991   0.299023   2.889  0.00441 **
   .
STATION12    0.055935   0.203888   0.274  0.78419
STATION13   -0.731586   0.172608  -4.238 3.85e-05 ***
STATION14    0.001491   0.299023   0.005  0.99603
STATION15    0.750722   0.172608   4.349 2.47e-05 ***
```