STAT 461: Lab 10 - ANCOVA

1 Continuous Factors

So far in this course we have assumed that our factors have a set number of discrete levels. For example, we might have 3 different fertilizers, or 2 different ovens. However, sometimes important factors are better represented as being continuous, rather than discrete.

1.1 Example: Flouride

Quade (1992) reports data obtained by Cartwrigtht, Lindahl and Bawden (1968) from an experiment to reduce dental cavities. The goal was to identify which flouride treatments were more effective than others at reducing the number of new cavities in children. As older children are in general more likely to get cavities than younger children, the age of each child in the study is an important factor that the researchers wanted to account for.

Two treatments, stannuous fluoride (SF), acid-phosphate fluoride (APF), and a placebo treatment of distilled water (W) were used in a two-year study. 69 female children completed the study. The number of new cavities (NumNewCavities) for each child was recorded, as well as the age of each child at the beginning of the study.

The data can be read into R by the following code:

```
library(lsmeans)
library(car)
library(multcompView)
library(lme4)
library(lmerTest)
options(contrasts = c("contr.sum", "contr.poly"))

flouride=read.table("flouride.csv",header=TRUE)
flouride
```

```
##    NewCavities age treatmt
## 1            4  13       W
## 2            4  17       W
## 3            4  16       W
## 4            1  13       W
## 5            4  10       W
## 6            3  17       W
## 7            4  13       W
## 8            2   9       W
## 9            2  14      SF
## 10           3  14      SF
## 11           3  11     APF
## 12           0  15     APF
## 13           2  11     APF
## 14           4   7     APF
## 15           4  11     APF
## 16           4  16       W
## 17           4  16       W
## 18           1   7       W
## 19           3  11       W
```

```
## 20           5  15       W
## 21           3  14      SF
## 22           2  11      SF
## 23           2   9      SF
## 24           1  17      SF
## 25           1  14      SF
## 26           0  13      SF
## 27           3   9      SF
## 28           2   9      SF
## 29           3  15      SF
## 30           2  10      SF
## 31           0  10     APF
## 32           0  15     APF
## 33           4  11     APF
## 34           0   9     APF
## 35           1   9     APF
## 36           0  16     APF
## 37           2   8       W
## 38           5  16       W
## 39           3  14       W
## 40           3  16       W
## 41           4  12       W
## 42           3   8       W
## 43           3  14       W
## 44           2   9      SF
## 45           4  15      SF
## 46           4  14      SF
## 47           1  13      SF
## 48           3  12      SF
## 49           3  12      SF
## 50           3  14      SF
## 51           2  13      SF
## 52           0  14      SF
## 53           3  14      SF
## 54           2  14     APF
## 55           4  11     APF
## 56          -1  14     APF
## 57           3   9     APF
## 58           2  11     APF
## 59           1  12     APF
## 60           1  14     APF
## 61           3  10     APF
## 62           0  12     APF
## 63           2  11     APF
## 64           2  11     APF
## 65           1  16     APF
## 66           2  15     APF
## 67           0  10     APF
## 68           1   6     APF
## 69           1   7     APF
```

Our response variable is `NewCavities`, and `treatmt` is clearly a fixed factor. The `age` of each child is also a factor, but as-is it has a LOT of levels in it. We could fit a 2-way crossed model to this data, by assuming

that each unique `age` is a factor level.

$$Y_{ijt} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijt}, \quad \epsilon_{ijt} \overset{iid}{\sim} N(0, \sigma^2)$$

where \$i = \${SF,APF,W} and $j = \{6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17\}$.

```
## fit model
fit.2way=aov(NewCavities~as.factor(age)+treatmt+as.factor(age):treatmt,data=flouride)

## ANOVA table
anova(fit.2way)
```

```
## Analysis of Variance Table
##
## Response: NewCavities
##                          Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(age)           11 12.270  1.1155  0.8392   0.60328
## treatmt                   2 43.577 21.7887 16.3915 5.874e-06 ***
## as.factor(age):treatmt 14 31.421  2.2443  1.6884   0.09627 .
## Residuals                41 54.500  1.3293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this analysis, we find no evidence that age affects the mean number of cavities in children. However, this analysis assumes that each age observed is its own factor. It also doesn't allow us to extrapolate to children of slightly different ages.

2. The analysis of covariance (ANCOVA)

The analysis of covariance (ANCOVA) is an ANOVA analysis where one or more factors are continuous in nature. Instead of treating each level of age as being a completely separate factor, we instead assume that the relationship between the factor x_j=(age of j-th child) and the mean response $E(Y_{ijt})$ is linear. To clarify this, let's remove `treatmt` from the analysis, and consider a model with one factor (x_{it}=age of j-th child):

$$Y_{it} = \mu + \beta * x_{it} + \epsilon_{it}, \quad \epsilon_{it} \overset{iid}{\sim} N(0, \sigma^2)$$

Under this model, the mean response is

$$E(Y_{it}) = \mu + \beta * x_{it}$$

which is a linear function of age (x_{it}), where μ is the intercept and β is the slope of the linear function. So if β is positive, then mean number of cavities increases with age, and if β is negative, then the mean number of cavities decreases with age.

The main benefits of this model are:

1. there is only 1 parameter to estimate for the effect of age, as β is the amount that average number of cavities increases when the age of a child increases by one year.
2. the linear relationship allows for the model to predict average number of cavities for ages that were not directly observed in the study.

A possible problem with this model is that

3. it assumes that the affect of age is linear. This could be relaxed by putting in quadratic or cubic terms when appropriate.

2.1 An ANCOVA model for the flouride data

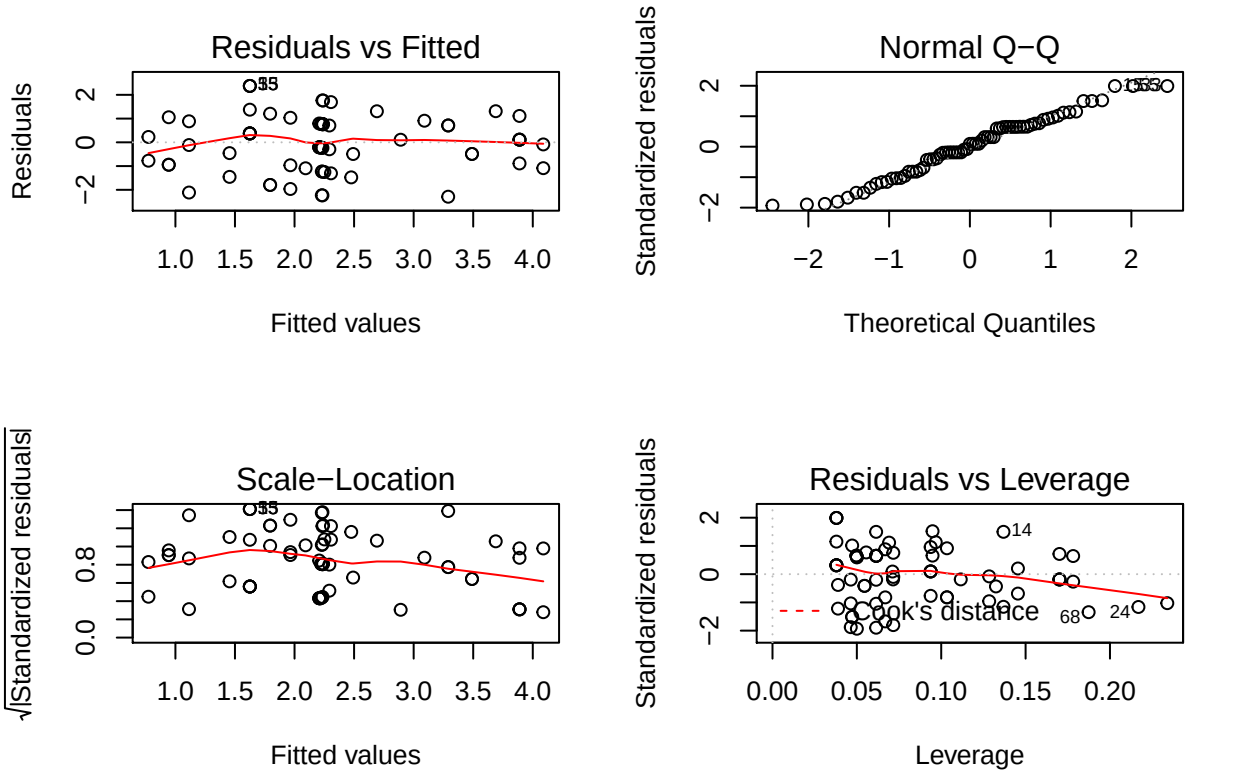
Now consider a model where `treatmt` is seen as a categorical factor, and `age` is seen as a continuous factor. We can still consider them to be crossed factors, so we include their interaction term:

$$Y_{it} = \mu + \alpha_i + \beta * x_j + (\alpha\beta)_i * x_{it} + \epsilon_{it}, \quad \epsilon_{ijt} \overset{iid}{\sim} N(0, \sigma^2)$$

This model is fit in R using the following code. Note that **the only difference is that we do NOT tell R to treat age as a factor**:

```
## fit model
fit.ANCOVA=aov(NewCavities~treatmt+age+age:treatmt,data=flouride)

## examine residuals
par(mfrow=c(2,2))
plot(fit.ANCOVA)
```




```
par(mfrow=c(1,1))

## ANOVA table
anova(fit.ANCOVA)
```

```
## Analysis of Variance Table
##
## Response: NewCavities
##             Df Sum Sq Mean Sq F value    Pr(>F)
## treatmt      2 35.038 17.5189 11.8461 4.294e-05 ***
## age          1  0.103  0.1026  0.0694   0.79307
## treatmt:age  2 13.459  6.7294  4.5504   0.01426 *
## Residuals   63 93.169  1.4789
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We first look at the interaction line. The "age:treatmt" line tests the null hypothesis that

$$H_0: \quad (\alpha\beta)_i = 0, \text{ for all } i$$

versus the alternative hypothesis at least one of the intercept terms $(\alpha\beta)_i$ is not zero. The "age" line tests the null hypothesis that the slope β is zero:

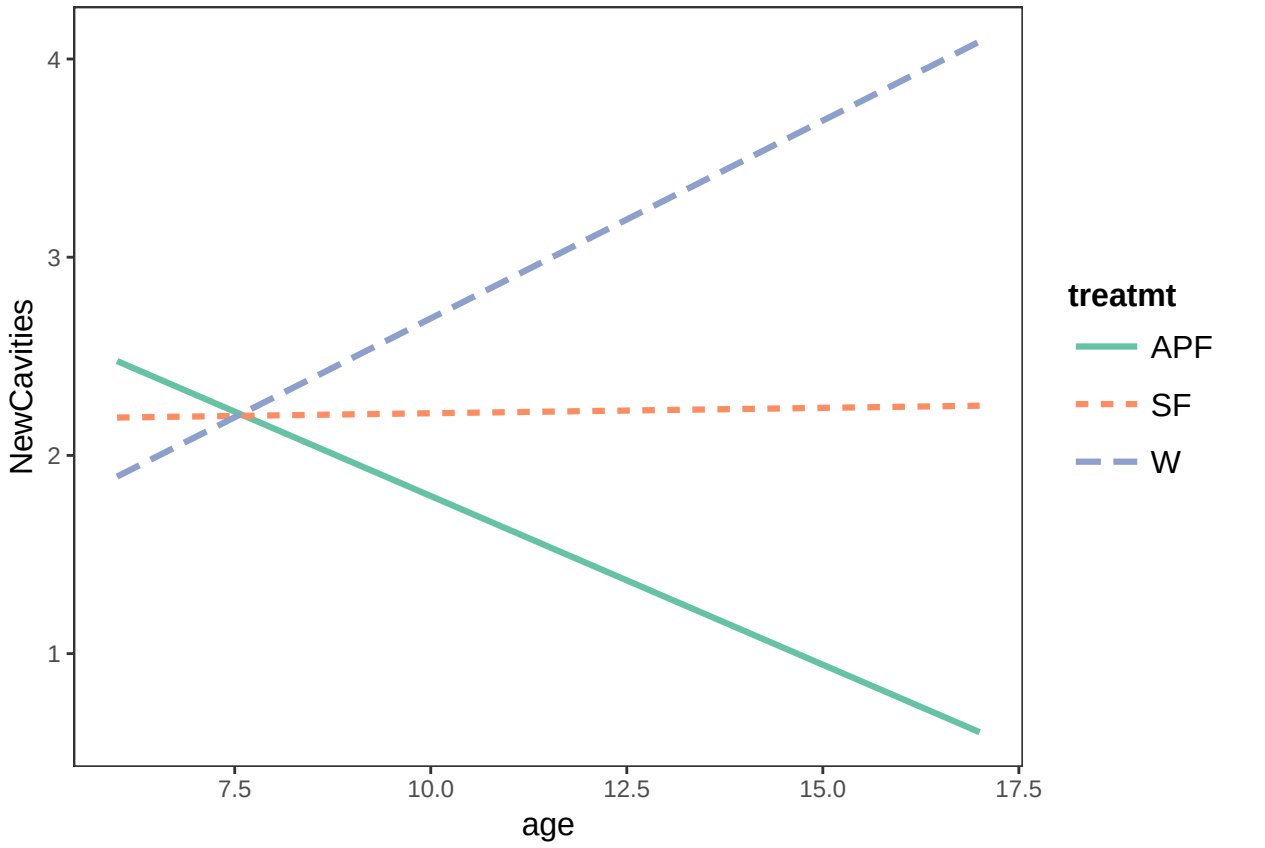
$$H_0: \quad \beta = 0$$

versus the alternative hypothesis that $\beta \neq 0$.

Hypothesis tests for categorical factors are identical to what we have done before, so the "treatmt" line is interpreted as like a 1-way ANOVA factor if the interaction term is found to be zero. In this case, the interaction term is significant, so we do NOT interpret the "treatmt" line by itself. Instead, we just look at the interactions.

When we had interactions of categorical factors, we would check for pairwise differences between the combinations of factor levels. In ANCOVA, we don't have factor levels for "age", so the best way to look for differences is to plot the estimated regression lines.

```
library(jtools)
interact_plot(fit.ANCOVA,pred="age",modx="treatmt")
```



Here we see that for young children, there is no difference between the average number of cavities, but as they age, the effect of flouride becomes very important, with children getting water having more cavities on average than children getting "SF", who get more cavities on average than children getting "APF".

Homework

1. Consider an experiment to determine whether or not having an automatic transmission affects gas mileage in cars. It is also known that the horsepower of a car affects gas mileage, so the horsepower of each car used is also recorded. The following R code will read in the data to R

```
cars <- mtcars[,c("am","mpg","hp")]
head(cars)
```

```
##                   am  mpg  hp
## Mazda RX4          1 21.0 110
## Mazda RX4 Wag      1 21.0 110
## Datsun 710         1 22.8  93
## Hornet 4 Drive     0 21.4 110
## Hornet Sportabout  0 18.7 175
## Valiant            0 18.1 105
```

```
## type "cars" into R to see the full data
```

Fit an appropriate model to this data, and interpret the results.

2. The "high school and beyond" survey looked at differences in socioeconomic status, race, gender, and types of schools and students' SAT scores. The data can be read into R using the following code:

```
hsb2= read.table("hsb2.csv")## from https://stats.idre.ucla.edu/stat/data/hsb2.csv
head(hsb2)
```

```
##   ses schtyp write math
## 1   1      1    52   41
## 2   2      1    59   53
## 3   3      1    33   54
## 4   3      1    44   47
## 5   2      1    52   57
## 6   2      1    52   51
```

```
## type "hsb2" into R to see the full data
```

The goal of this study is to see if socio-economic status (ses) and school type (schtyp) affect students math scores, after accounting for how well they do on their writing score. Student's writing score should be treated as a continuous covariate. Conduct a full analysis of this data. Fit an appropriate model, and interpret the results.