Analysis of Variance (Unbalanced Case)

Review: Balanced Case

- Breaking up variation of data based on source
- Terms **orthogonal** i.e., You can uniquely decompose variation
- The **order of variables** does not change the result

This is not the case for an unbalanced case. The order of the variables does change the result.

Review: Balanced Case

This code only focuses on the two main effects, does not include the interaction effect)

```
summary(aov(Toothlength ~ Supplement + Dose, data=tooth))

##             Df Sum Sq Mean Sq F value   Pr(>F)
## Supplement   1  205.4   205.4   14.02 0.000429 ***
## Dose         2 2426.4  1213.2   82.81  < 2e-16 ***
## Residuals   56  820.4    14.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

summary(aov(Toothlength ~ Dose + Supplement, data=tooth))

##             Df Sum Sq Mean Sq F value   Pr(>F)
## Dose         2 2426.4  1213.2   82.81  < 2e-16 ***
## Supplement   1  205.4   205.4   14.02 0.000429 ***
## Residuals   56  820.4    14.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
```

Example: ozkids data

Continuous response:

Testing to see the significant effect of predictors on days absent.

- **days**: days absent

Categorical predictors:

- **origin**: Aboriginal or not
- **sex**: male or female
- **grade**: level in school
- **type**: type of learner

```
str(ozkid)
```

```
## 'data.frame':    154 obs. of  5 variables:
##  $ origin: Factor w/ 2 levels "A","N": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex   : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ grade : Factor w/ 4 levels "F0","F1","F2",..: 1 1 1 1 1 1 1 1 2 2
..
##  $ type  : Factor w/ 2 levels "AL","SL": 2 2 2 1 1 1 1 1 2 2 ...
##  $ days  : int  2 11 14 5 5 13 20 22 6 6 ...
```

Unbalanced Case

- Still want to decompose variation – to test significance of variables
- Effects are not orthogonal, i.e., Decomposition is not unique

There is some overlap

```
table(ozkid$origin); table(ozkid$sex)  # check unbalance
```

```
##
##  A  N
## 74 80
```

```
##
##  F  M
## 84 70
```

Confirmed that the data is unbalanced.

Unbalanced Case

```
summary(aov(days ~ origin + grade, data=ozkid))

##              Df Sum Sq Mean Sq F value   Pr(>F)
## origin        1   2646  2645.7  11.580 0.000856 ***
## grade         3   2020   673.4   2.947 0.034821 *
## Residuals   149  34040   228.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

==================================================================
summary(aov(days ~  grade + origin, data=ozkid))

##              Df Sum Sq Mean Sq F value  Pr(>F)
## grade         3   2277   759.1   3.323 0.02149 *
## origin        1   2389  2388.5  10.455 0.00151 **
## Residuals   149  34040   228.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When you change the order of the variables, it completely changes the results when comparing the top to the bottom code

Types of Sums of Squares

- 3-way ANOVA model with variables A, B, and C
- Notations:
- **SS (C|A B)**-additional contribution when C is added to model containing A and B
- **SS(C)** ≠ **SS(C|AB)**if unbalanced
- **SS(C)** = **SS(C|AB)**if balanced

"C" given (|) "A" and "B"

"Types" are used the calculate the sum of squares when you have an unbalanced case.

Type I

- Sequential sum of squares (SS)
- Results from aov() or lm()
- **Additional variation explained by the model when that term is added to terms already in**
  - If we run aov(Y~A+B+C), it displays SS(A), SS(B|A), and SS(C|A,B)

Sequential means: "A" goes in, then when "B" goes in, it is added to "A", and when "C" goes in, it is added to "A" and "B".

  - If we run aov(Y~B+A+C), it displays SS(B), SS(A|B), and SS(C|A,B)
- Order does matter, like which variable is added in the model first

With Type 1, order matters, but the order will not impact the Total SS

Type 3 enables you to see the unique contribution of each variable

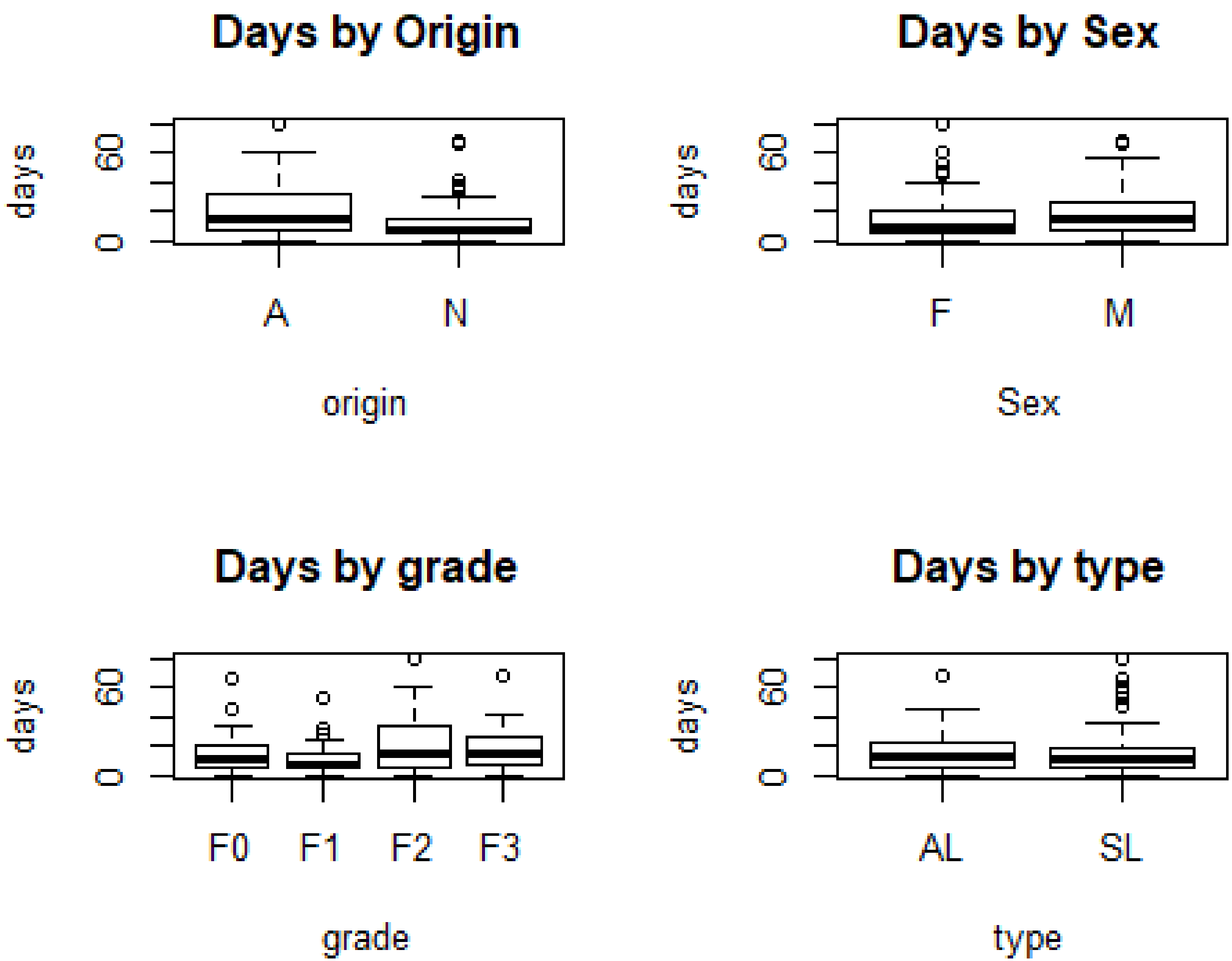
Type III

- Partial sums of squares
- Available from Anova() in package “car”
- **Explained variation that term adds when all other terms are already included**
  - If we run aov(Y~A+B+C), it displays SS(A|B,C), SS(B|A,C), and SS(C|A,B)
  - If we run aov(Y~B+A+C), it displays SS(B|A,C), SS(A|B,C), and SS(C|A,B)
- Order does not matter. Consistent outputs
- https://www.r-bloggers.com/anova-%E2%80%93-type-iiiiii-ss-explained/

Exercise: data exploration

- Boxplots of days by Origin
- Or by sex, grade, or type
- Informally compare means and variations of absent day among groups

Exercise: data exploration



Example: **origin** and **grade** Models

- Days absent with **origin** and **grade** predictors
- See Type I and Type III sums of squares
- Reverse order of terms (e.g. **grade** first and then **origin**)
- What stays the same, and what changes?
- Impact of reversing main effects on interaction model?

Example: **origin** and **grade** Models

```
aov.res1= aov(days ~  grade + origin, data=ozkid) # type 1 test

Anova(aov.res1, type=3)  # type 3 test

## Anova Table (Type III tests)
##
## Response: days
##             Sum Sq  Df F value    Pr(>F)
## (Intercept)   8148   1 35.6655 1.648e-08 ***
## grade         2020   3  2.9474  0.034821 *
## origin        2389   1 10.4550  0.001506 **
## Residuals    34040 149
====================================================================

aov.res2= aov(days ~  origin + grade, data=ozkid)

Anova(aov.res2, type=3)

## Anova Table (Type III tests)
##
## Response: days
##             Sum Sq  Df F value    Pr(>F)
## (Intercept)   8148   1 35.6655 1.648e-08 ***
## origin        2389   1 10.4550  0.001506 **
## grade         2020   3  2.9474  0.034821 *
## Residuals    34040 149
```

Exercise: Type III Analysis in Four-Way Main Effects Model

- Type III SS for the four-way main effects model
- Backward elimination and conclusions about main effects to keep in the model?
- Get Type I SS for each of the orderings of the terms we might want to keep
- Could we further reduce the main effects we would want to keep in the model?

Exercise: Multiple Comparisons

- Fit model with previous main effects and all interactions between them.
- Which main effects and interactions kept?
- Do Tukey multiple comparison for the main effects and note differences.
- Significantly different groups?