Discussion6

Jing Lyu

2/15/2023

Two-way ANOVA with fixed effects

In this section, we'll use the built-in R data set 'ToothGrowth'. It includes information from a study on the effects of vitamin C on tooth growth in Guinea pigs.

The trial used 60 pigs who were given one of three vitamin C dose levels (0.5, 1, or 2 mg/day) via one of two administration routes: orange juice (OJ) or ascorbic acid (VC).

```
library(dplyr)
dat = ToothGrowth
str(dat)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

R treats 'dose' as a numeric variable based on the output. We'll transform it to a factor variable.

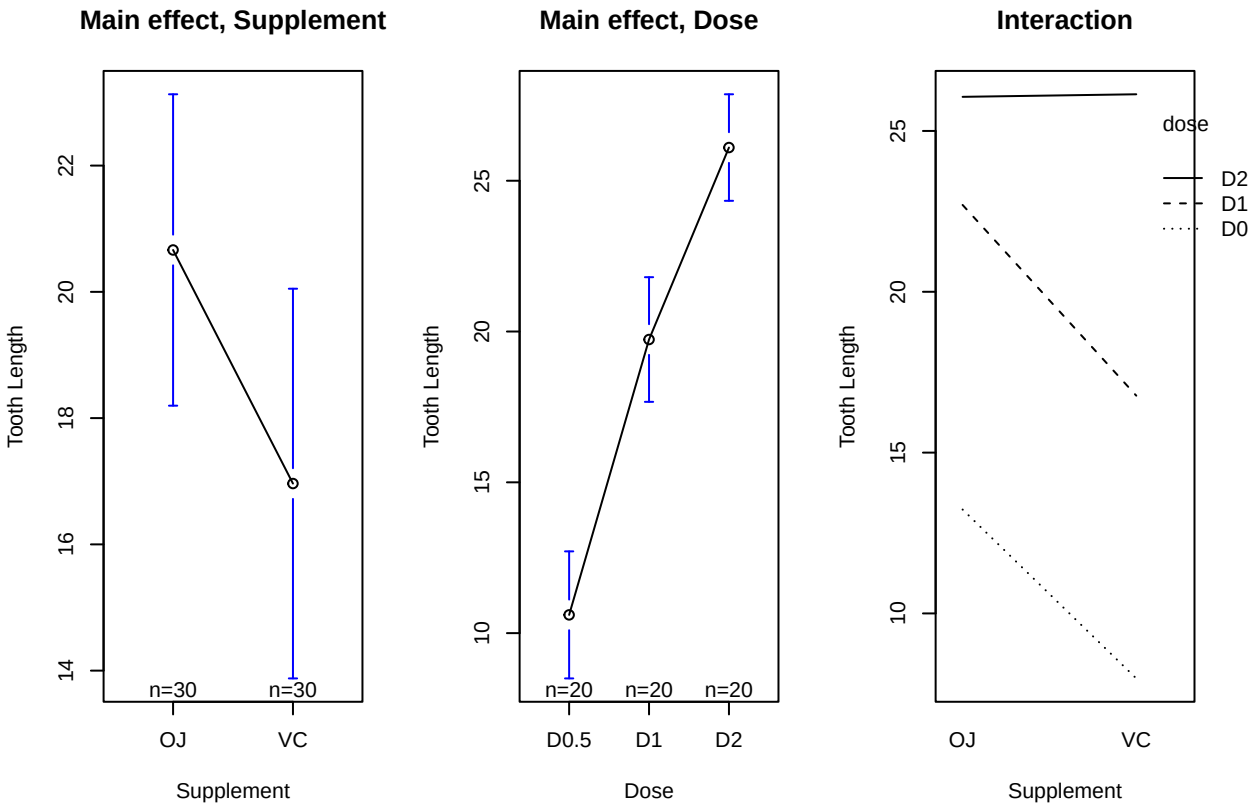
```
dat$dose <- factor(dat$dose,
                   levels = c(0.5, 1, 2),
                   labels = c("D0.5", "D1", "D2"))
table(dat$supp,dat$dose)
```

```
##
##      D0.5 D1 D2
##   OJ   10 10 10
##   VC   10 10 10
```

We have a well-balanced design.

Main effects plots and the interaction plot

```
library(gplots)
par(mfrow=c(1,3))
plotmeans(len~supp,data=dat,xlab="Supplement",ylab="Tooth Length", main="Main effect, Supplement")
plotmeans(len~dose,data=dat,xlab="Dose",ylab="Tooth Length", main="Main effect, Dose")
dose=dat$dose
supp=dat$supp
len=dat$len
interaction.plot(supp,dose,len,xlab="Supplement",ylab="Tooth Length", main="Interaction")
```




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

There seems to be some interaction effects.

```
model2 <- aov(len ~ supp * dose, data = dat)
summary(model2)
```

Modeling

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## supp          1  205.4   205.4  15.572 0.000231 ***
## dose          2 2426.4  1213.2  92.000  < 2e-16 ***
## supp:dose     2  108.3    54.2   4.107 0.021860 *
## Residuals    54  712.1    13.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction has p-value smaller than 0.05, indicating the relationship between dose and tooth length is significantly influenced by the supplement technique (or you can say the relationship between supplement technique and tooth length is influenced by the dose).

The most important factor variable is `dose`. We can conclude that modifying the delivery technique (supp) or the vitamin C dose will have a major impact on the mean tooth length.

Multiple pairwise comparisons

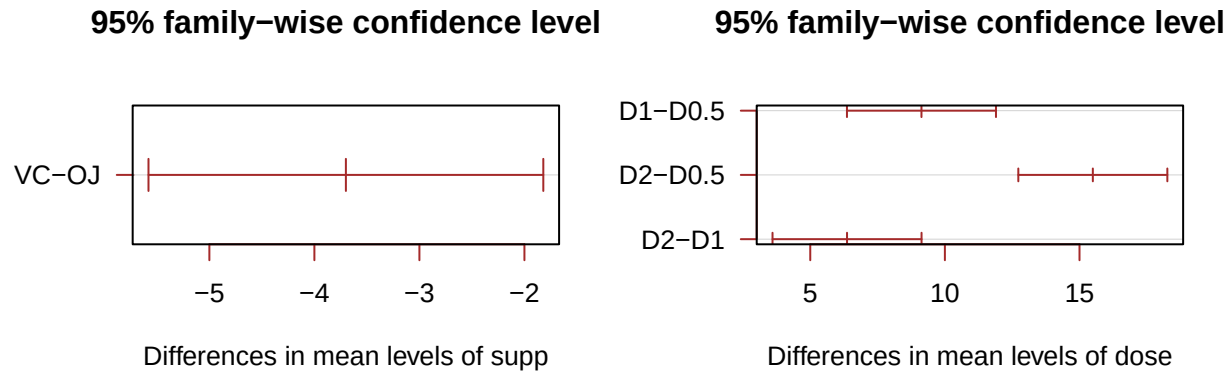
Significant p-values in an ANOVA test shows that some of the group means differ, but we don't know which pairs of groups have different means.

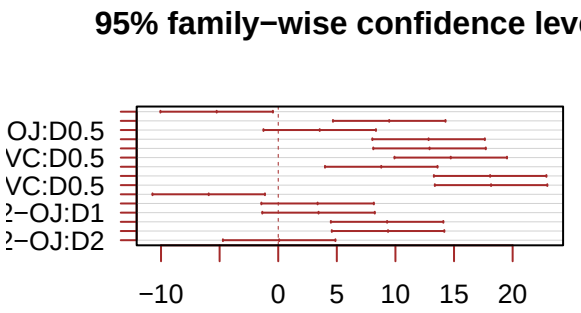
We may want to explore what the best combination of supplement and dose is, or whether such combination

exists. The "best combination" represents the cell with the highest cell mean compared to other cells. To do all pairwise comparisons, we need Tukey-Kramer method.

- Tukey's method

```
T.ci=TukeyHSD(model2,conf.level = 0.95)
par(mfrow=c(2,2))
plot(T.ci, las=1 , col="brown")
```





Simultaneous confidence intervals of factor `dose`:

```
# Only show all pairwise comparisons in factor dose
TukeyHSD(model2, which = "dose")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = len ~ supp * dose, data = dat)
##
## $dose
##           diff       lwr       upr   p adj
## D1-D0.5  9.130  6.362488 11.897512 0.0e+00
## D2-D0.5 15.495 12.727488 18.262512 0.0e+00
## D2-D1    6.365  3.597488  9.132512 2.7e-06
```

The output shows that all pairwise comparisons in factor `dose` with an adjusted p-value of 0.05 are significant.

To find the best combination, we only need to focus on the differences of the two largest means.

```
idx=list();
idx[[1]]=dat$supp;idx[[2]]=dat$dose;
(means.comb=tapply(dat$len, INDEX=idx,mean))
```

```
##     D0.5    D1    D2
## OJ 13.23 22.70 26.06
## VC  7.98 16.77 26.14
```

From this table, the two cells are (OJ,D2) and (VC,D2).

Then, let's find the confidence interval corresponding to the difference between (OJ,D2) and (VC,D2).

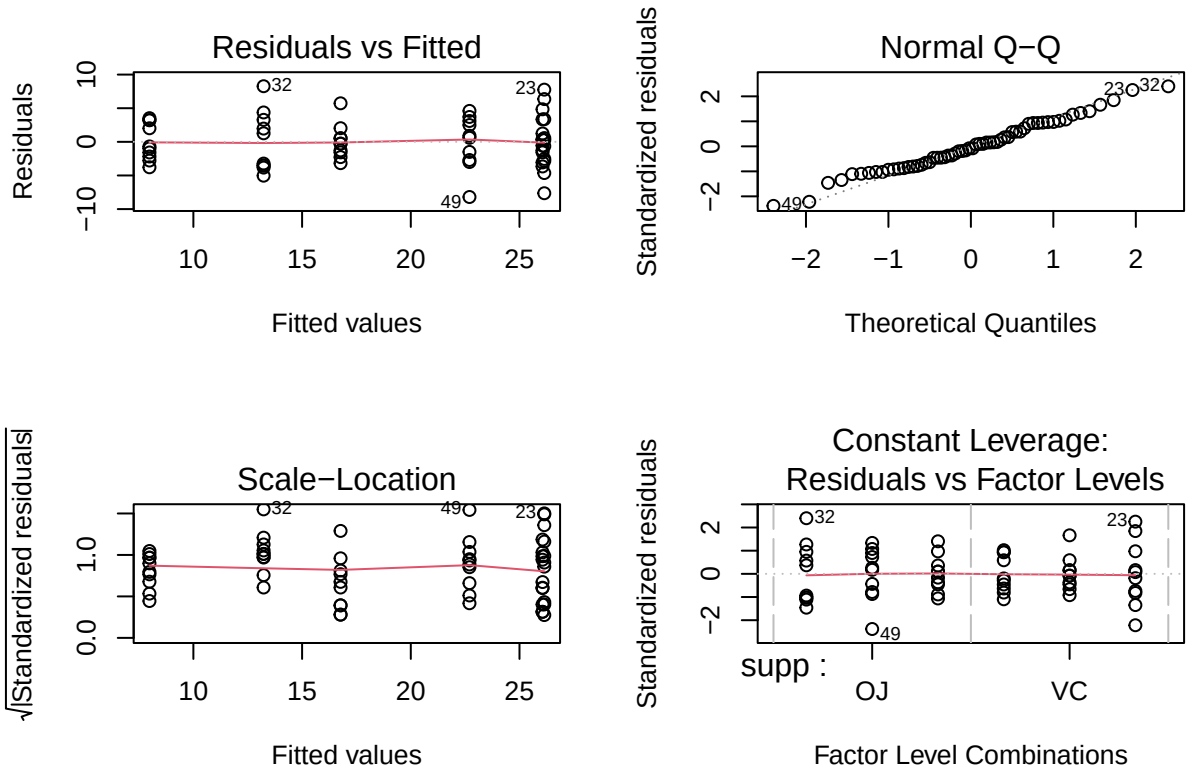
```
T.ci[['supp:dose']]['VC:D2-OJ:D2',]
```

```
##      diff       lwr       upr     p adj
##  0.080000 -4.718124  4.878124  1.000000
```

Since the p-value is not significant, we can conclude there is no enough evidence of distinguishing (OJ,D2) and (VC,D2). We can not decide which one is the best combination.

Diagnostics

```
par(mfrow=c(2,2))
plot(model2)
```



It seems points 32, 23 and 49 are outliers, which can have a significant impact on normality and variance homogeneity. It may be beneficial to remove outliers.

Check the homogeneity of variances with Levene's test:

```
library(car)
leveneTest(len ~ supp*dose, data = dat)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  5  1.7086 0.1484
##       54
```

Assuming significance level is 0.05, we can infer that the variations in the different treatment groups are homogeneous.

Check the normality with Shapiro-Wilk test:

```
model2.residuals = residuals(model2)
shapiro.test(model2.residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2.residuals
## W = 0.98499, p-value = 0.6694
```

Normality assumption is satisfied.

Type I, II and III ANOVA

There are three methods for splitting the total variation of a dependent variable: Type I, Type II and Type III Sums of Squares. They do not give the same result in case of unbalanced data.

Suppose the model has two independent variables A and B.

1. Type I ANOVA

Type I Sums of Squares are Sequential, so the order of variables in the models makes a difference. Sums of Squares are Mathematically defined as:

- SS(A) for independent variable A
- SS(B|A) for independent variable B
- SS(AB|A,B) for interaction effect

2. Type II ANOVA

Type II Sums of Squares are not sequential. It should be used if there is no interaction effect. Sums of Squares are Mathematically defined as:

- SS(A | B) for independent variable A
- SS(B | A) for independent variable B

3. Type III ANOVA

Type III Sums of Squares are not sequential. It's also called partial sums of squares. Sums of Squares are Mathematically defined as:

- SS(A | B, AB) for independent variable A
- SS(B | A, AB) for independent variable B

Example: For this section, we use `Wage` dataset. We are interested in investigating the relationship between wages and two demographic factors: ethnicity and occupation.

It is an unbalanced design.

```
wage = read.csv('Wage.csv')
wage$ethnicity = as.factor(wage$ethnicity)
wage$occupation = as.factor(wage$occupation)
table(wage$ethnicity, wage$occupation)
```

```
##
##            management office sales services technical worker
##   cauc             46     77    34       60        93    130
##   hispanic          3      5     1        6         5      7
##   other             6     15     3       17         7     19
```

aov performs type I ANOVA. Different orders generate different test results for the main effects. But the inference for the interaction term is the same. It appears the interaction effects are not significant.

```
fit1 = aov(wage ~ ethnicity + occupation + ethnicity:occupation, data=wage)
summary(fit1)
```

```
##                       Df Sum Sq Mean Sq F value Pr(>F)
## ethnicity              2    173    86.3   3.986 0.0191 *
## occupation             5   2459   491.7  22.704 <2e-16 ***
## ethnicity:occupation  10    270    27.0   1.247 0.2579
## Residuals            516  11175    21.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit2 = aov(wage ~ occupation + ethnicity + ethnicity:occupation, data=wage)
summary(fit2)
```

```
##                       Df Sum Sq Mean Sq F value Pr(>F)
## occupation             5   2538   507.5  23.435 <2e-16 ***
## ethnicity              2     94    46.8   2.159  0.116
## occupation:ethnicity  10    270    27.0   1.247  0.258
## Residuals            516  11175    21.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Type I ANOVA is rarely what we want in practice. The variable is considered as the most important is just because it was specified first in the model.

Anova function from car package can generate type II and III ANOVA.

```
Anova(lm(wage ~ ethnicity + occupation, data=wage), type = 'II')
```

```
## Anova Table (Type II tests)
##
## Response: wage
##             Sum Sq  Df F value Pr(>F)
## ethnicity     93.5   2  2.1491 0.1176
## occupation  2458.6   5 22.5977 <2e-16 ***
## Residuals  11445.5 526
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(lm(wage ~ 0 + ethnicity * occupation, data=wage), type = 'III')
```

```
## Anova Table (Type III tests)
##
## Response: wage
##                       Sum Sq  Df  F value Pr(>F)
## ethnicity             8965.1   3 137.9818 <2e-16 ***
## occupation            2430.0   5  22.4402 <2e-16 ***
## ethnicity:occupation   270.2  10   1.2474 0.2579
## Residuals            11175.3 516
```

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

Another way to test interaction term:

```
full.model = aov(wage ~ ethnicity * occupation, data=wage)
reduced.model = aov(wage ~ ethnicity +  occupation, data=wage)
anova(reduced.model, full.model)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ ethnicity + occupation
## Model 2: wage ~ ethnicity * occupation
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    526 11446
## 2    516 11175 10    270.16 1.2474 0.2579
```

I would suggest to test for the interaction term first (SS(AB|A,B)) and only if the interaction is not significant, continue with the analysis for main effects.

If there is indeed no interaction, then type II is statistically more powerful than type III.