Experimental design and ANOVA examples

AUTHOR
Alex Sanchez-Pla

PUBLISHED
February 14, 2023

Example 1. A completely randomized experiment

Description

Gene therapy experiment: compare four techniques for correcting faulty genes

- A: normal gene inserted in non-specific location
  - B: Abnormal gene swapped for a normal gene
  - C: Abnormal gene repaired through selective reverse mutation
  - D: Regulation of a particular gene altered

Randomization

20 independent individuals are selected.

```
mice <- paste0("m",1:20)
```

Treatments are assigned at random.

```
randomized<- sample (mice, 20)
TREAT<-rep(LETTERS[1:4], each=5)
names(randomized) <- TREAT
```

Data collection

Once the randomization is done, the experiment is performed and gene expression is measured on each mouse.

```
RESP <- c(96,99,100,104,84,91,90,75,80,90,70,90,84,76,78,78,87,67,66,76)
names(RESP) <- paste(names(randomized), randomized, sep=".")
```

Data Analysis

```r
dades<-data.frame (TREAT, RESP)
dades$TREAT <- as.factor(TREAT)

kableExtra::kable(dades) %>%  kableExtra::kable_styling(full_width=FALSE)
```

	TREAT	RESP
A.m14	A	96
A.m4	A	99
A.m9	A	100
A.m19	A	104
A.m5	A	84
B.m12	B	91
B.m17	B	90
B.m18	B	75
B.m15	B	80
B.m7	B	90
C.m13	C	70
C.m8	C	90
C.m11	C	84
C.m6	C	76
C.m20	C	78
D.m3	D	78
D.m2	D	87
D.m1	D	67
D.m10	D	66
D.m16	D	76

```r
model <- RESP ~ TREAT

aov1<-aov (model, data=dades)
summary(aov1)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
TREAT        3   1321   440.3   7.218 0.0028 **
Residuals   16    976    61.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model.tables(aov1)
```

```
Tables of effects

 TREAT
TREAT
    A     B     C     D
12.55  1.15 -4.45 -9.25
```

```r
model.tables(aov1,type="means") # Mitjanes dels grups
```

```
Tables of means
Grand mean

84.05

 TREAT
TREAT
   A    B    C    D
96.6 85.2 79.6 74.8
```

```r
(hsd=TukeyHSD(aov1,which="TREAT",conf=0.95))
```

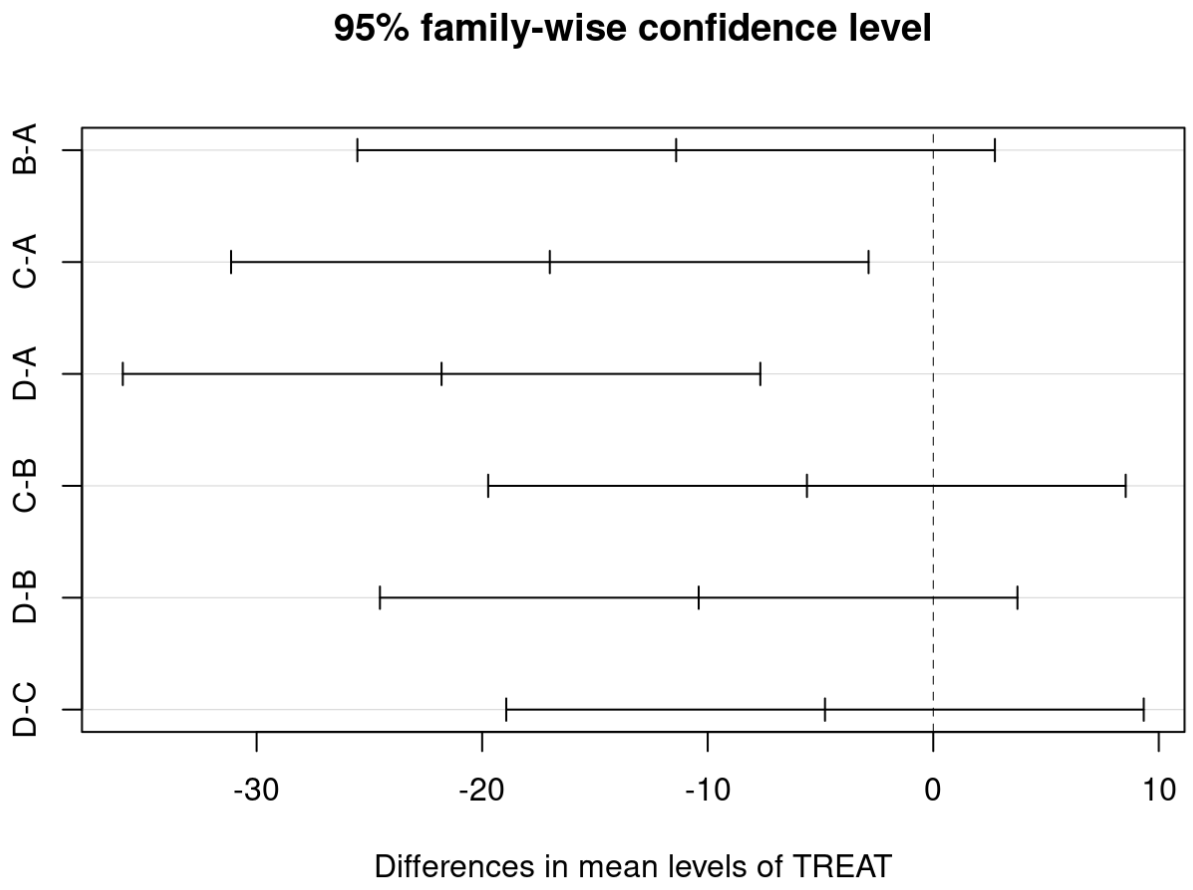
```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = model, data = dades)

$TREAT
     diff       lwr       upr     p adj
B-A -11.4 -25.5324  2.732395 0.1376990
C-A -17.0 -31.1324 -2.867605 0.0159107
D-A -21.8 -35.9324 -7.667605 0.0022116
C-B  -5.6 -19.7324  8.532395 0.6748712
```

```
D-B -10.4 -24.5324  3.732395 0.1933974
D-C  -4.8 -18.9324  9.332395 0.7670929
```

```
plot(hsd)
```



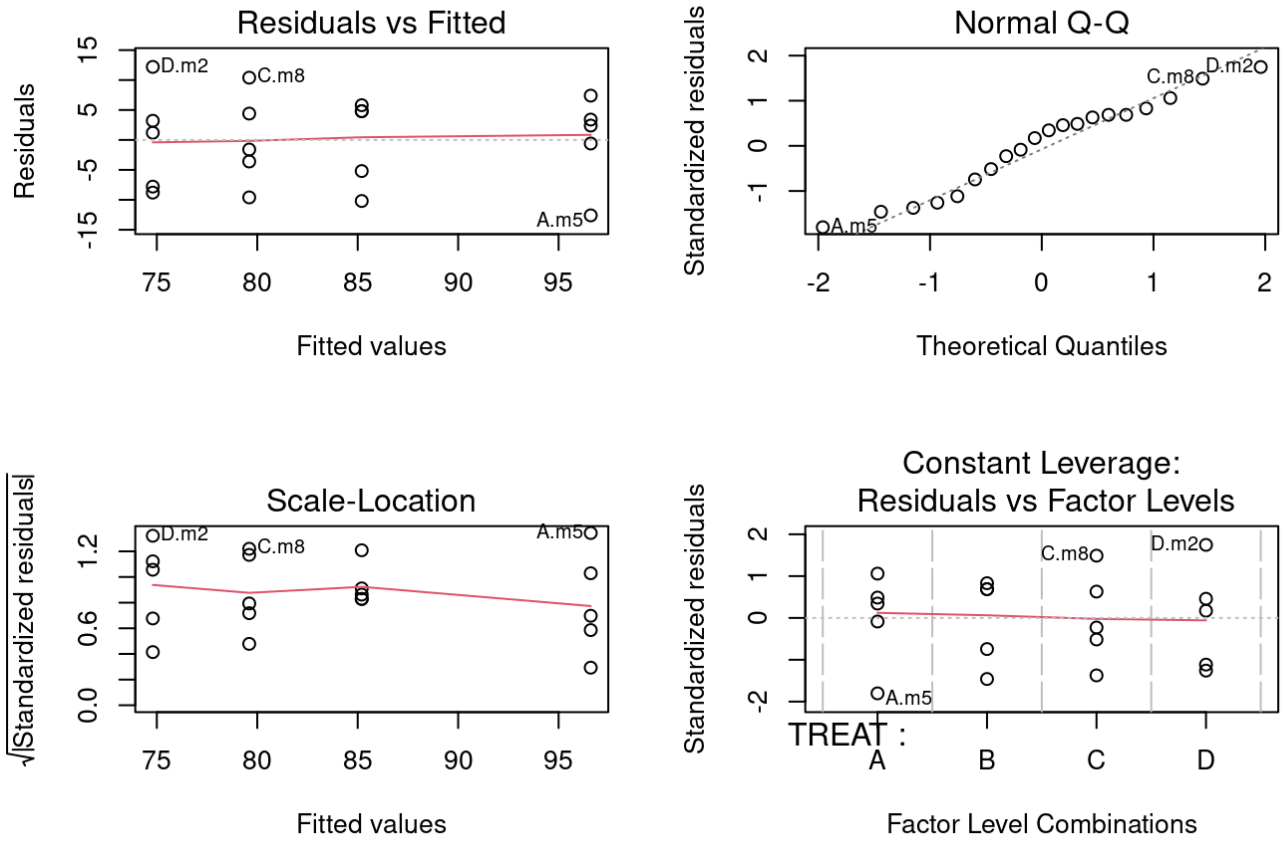
Model assumptions verification

ANOVA does not work under all circumstances. Some assumptions have to be true for the data:

- Homocedasticity or variance homogeneity
- Independence of errors
- Normality of errors

This can be a long process but plots are helpful to provide an overview of assumptions verification.

```
opt<- par(mfrow=c(2,2))
plot(aov1)
```




```
par(opt)
```

Exercises

A gene is suspected to have some connection with blood cancer. There are four stages of blood cancer: stage I, stage II, stage III, and stage IV.

For treating a patient, identification of blood cancer is crucial in the first three stages.

Three mRNA samples were collected from stage I, stage II, and stage III, respectively.

The experiment is repeated six times, as shown in the table.

Find whether there is any difference in mean expression values in the three mRNA samples.

Table 9.1.3 mRNA Samples, One-Way Classification

	1	2	3	4	5	6
mRNA – 1	95	98	100	105	85	88
mRNA – 2	94	92	78	88	92	91
mRNA – 3	72	88	82	73	75	77

Example 2: Randomized Block design

Randomization

Randomization is not illustrated here. Try to find out how you can do the randomization per blocks.

```
dades<-expand.grid(medium=1:4, Tx =1:6)
```

Data collection

Once treatments have been randomly assigned to each culture medium the experiment is performed and the data is collected.

```
regeneration <- c(34.98,41.22,36.94,39.97,40.89,46.69, 46.65, 41.90, 42.07, 49
dades <-cbind(dades, regeneration)
dades$medium<-as.factor(dades$medium)
dades$Tx<-as.factor(dades$Tx)
kableExtra::kable(dades) %>%  kableExtra::kable_styling(full_width=FALSE)
```

medium	Tx	regeneration
1	1	34.98
2	1	41.22
3	1	36.94
4	1	39.97
1	2	40.89

medium	Tx	regeneration
2	2	46.69
3	2	46.65
4	2	41.90
1	3	42.07
2	3	49.42
3	3	52.68
4	3	42.91
1	4	37.18
2	4	45.85
3	4	40.23
4	4	39.20
1	5	37.99
2	5	41.99
3	5	37.61
4	5	40.45
1	6	34.89
2	6	50.15
3	6	44.57
4	6	43.29

Data analysis

The analysis is based on the following linear model:

```
model.1<- regeneration ~ medium + Tx
model.1
```

```
regeneration ~ medium + Tx
```

The model is used to establish if there are any significant differences between treatments.

The differences between mediums are not relevant, because we assume they are different (that's why we block).

```
aov.1<-aov (model.1, data=dades)
summary(aov.1)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
medium       3  197.0   65.67   9.120 0.00112 **
Tx           5  201.3   40.26   5.592 0.00419 **
Residuals   15  108.0    7.20
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# equatiomatic::extract_eq(aov.1)
```

If we had ignored blocking, the result would have been different:

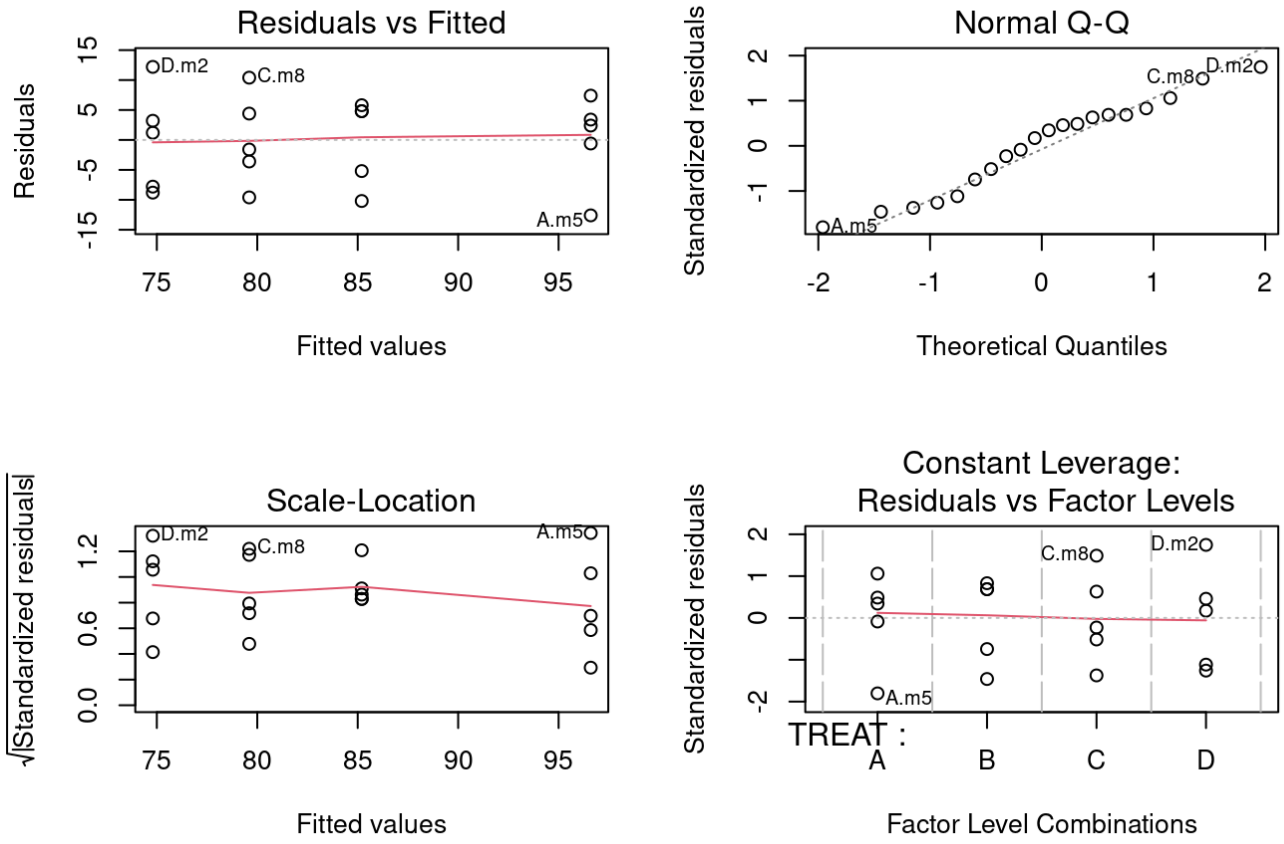
```
model.0<- regeneration ~  Tx
aov.0<-aov (model.0, data=dades)
summary(aov.0)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Tx           5  201.3   40.26   2.376 0.0802 .
Residuals   18  305.0   16.95
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Assumptions

Assumptions should always be checked when using ANOVA

```
opt<- par(mfrow=c(2,2))
plot(aov1)
```




```
par(opt)
```

Exercise

Effect of Atorvastatin (Lipitor) on Gene Expression in People with Vascular Disease (National Institutes of Health)

It has been known that atherosclerosis and its consequences (coronary heart disease and stroke) are the principal causes of mortality. Gene expression profiling of peripheral white blood cells provides information that may be predictive about vascular risk.

Table 9.2.3 gives gene expression meassurements, classified according to age group and dose level of Atorvastatin treatment.

Test whether the dose level and age groups significantly affect the gene expression.

Table 9.2.3 Gene Expression Data for Atherosclerosis

		Dose Level			
		1	2	3	4
Age Group	1	100	99	87	98
	2	95	94	83	92
	3	102	80	86	85
	4	84	82	80	83
	5	77	78	90	76
	6	90	76	74	85

Example 3: Factorial designs

- A study was conducted to study the effect of a drug and a diet on systolic blood pressure.
- 20 people with high blood pressure were randomized to one of four treatment conditions.
    - Control group (neither diet nor drug modification)
    - Diet modification only
    - Drug only
    - Modification of both drugs and diet
- At the end of the treatment period, systolic blood pressure was assessed.
- It is a factorial design in which each of the two treatments (drug, diet) can be randomly assigned to each individual.
- By having 20 individuals, there can be replicates of each treatment combination.

Randomization

We start by considering the possible combinations between all levels of Diet and Treatment.

```
dades<-expand.grid(replica=1:5, Diet = c("Standard", "Modified"), Treatment=c(
kableExtra::kable(dades) %>% kableExtra::kable_styling(full_width=FALSE)
```

replica	Diet	Treatment
1	Standard	Placebo
2	Standard	Placebo
3	Standard	Placebo
4	Standard	Placebo
5	Standard	Placebo
1	Modified	Placebo
2	Modified	Placebo
3	Modified	Placebo
4	Modified	Placebo
5	Modified	Placebo
1	Standard	Drug
2	Standard	Drug
3	Standard	Drug
4	Standard	Drug
5	Standard	Drug
1	Modified	Drug
2	Modified	Drug
3	Modified	Drug
4	Modified	Drug
5	Modified	Drug

Participants in the study (call them i1, i2, …i20) are randomly assigned to one of the combinations.

```
participants <- paste0("i", 1:20)
assignments <- cbind(dades, Partic=sample(participants))
```

```
kableExtra::kable(assignments) %>% kableExtra::kable_styling(full_width=FALSE)
```

replica	Diet	Treatment	Partic
1	Standard	Placebo	i9
2	Standard	Placebo	i18
3	Standard	Placebo	i3
4	Standard	Placebo	i6
5	Standard	Placebo	i11
1	Modified	Placebo	i15
2	Modified	Placebo	i2
3	Modified	Placebo	i7
4	Modified	Placebo	i14
5	Modified	Placebo	i12
1	Standard	Drug	i1
2	Standard	Drug	i10
3	Standard	Drug	i13
4	Standard	Drug	i19
5	Standard	Drug	i4
1	Modified	Drug	i17
2	Modified	Drug	i16
3	Modified	Drug	i5
4	Modified	Drug	i8
5	Modified	Drug	i20

Data collection

Next, the experiment is done and the data can be collected

```
WBC<-c( 2, .7, 1, 1.2, 1.3, 1.9, 1.9, 3.5, 1.2, 2.3, 2.4, 2.6, 1.9, 1.6, 1.7,
names(WBC)<- assignments$Partic
dades <- cbind(assignments, WBC)
kableExtra::kable(dades) %>% kableExtra::kable_styling(full_width=FALSE)
```

	replica	Diet	Treatment	Partic	WBC
i9	1	Standard	Placebo	i9	2.0
i18	2	Standard	Placebo	i18	0.7
i3	3	Standard	Placebo	i3	1.0
i6	4	Standard	Placebo	i6	1.2
i11	5	Standard	Placebo	i11	1.3
i15	1	Modified	Placebo	i15	1.9
i2	2	Modified	Placebo	i2	1.9
i7	3	Modified	Placebo	i7	3.5
i14	4	Modified	Placebo	i14	1.2
i12	5	Modified	Placebo	i12	2.3
i1	1	Standard	Drug	i1	2.4
i10	2	Standard	Drug	i10	2.6
i13	3	Standard	Drug	i13	1.9
i19	4	Standard	Drug	i19	1.6
i4	5	Standard	Drug	i4	1.7
i17	1	Modified	Drug	i17	0.4
i16	2	Modified	Drug	i16	0.2
i5	3	Modified	Drug	i5	0.1
i8	4	Modified	Drug	i8	0.4
i20	5	Modified	Drug	i20	0.3

Modeling and Analysis

First, consider a model that ignores interaction between treatment and diet.

```
model.1<- WBC ~ Diet + Treatment # this model ignores interaction
aov.1<-aov (model.1, data=dades)
summary(aov.1)
```

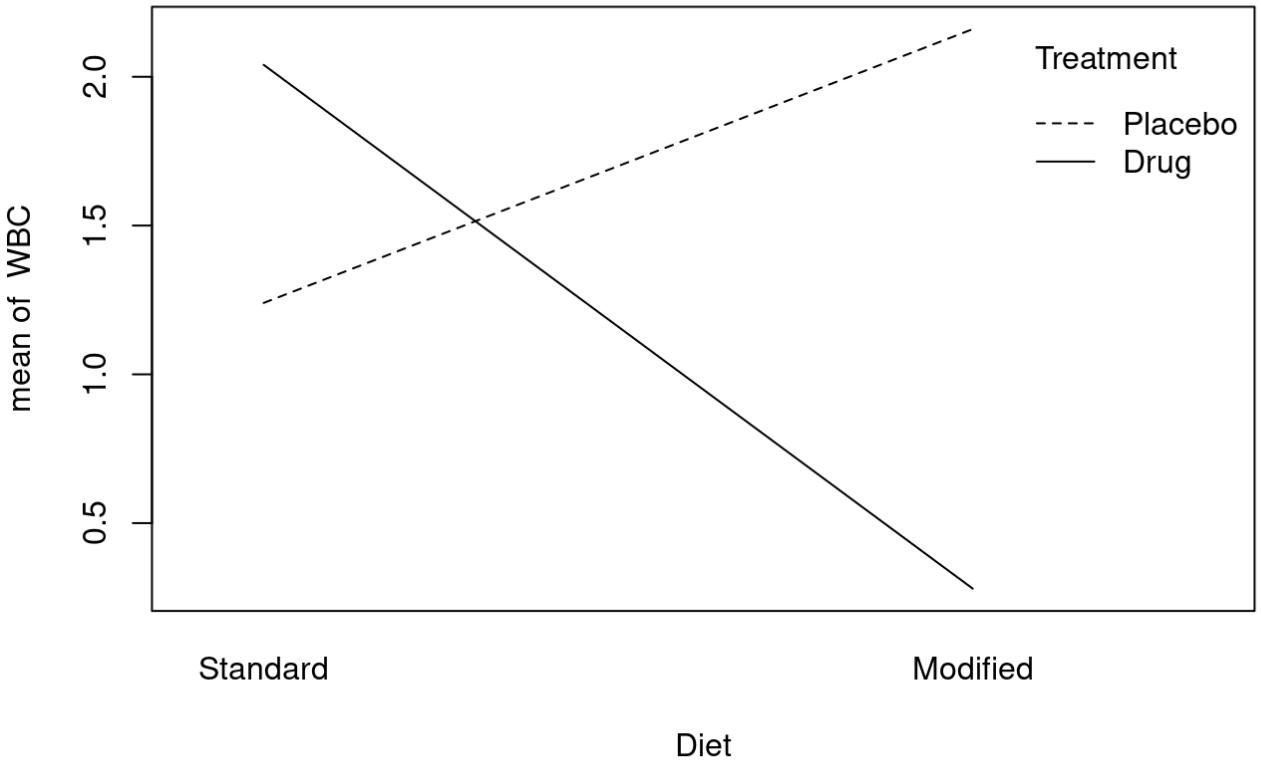
```
            Df Sum Sq Mean Sq F value Pr(>F)
Diet         1  0.882  0.8820   1.101  0.309
Treatment    1  1.458  1.4580   1.820  0.195
Residuals   17 13.622  0.8013
```

Now add an interaction between Diet and Treatment.

```
model.2<- WBC ~ Diet + Treatment + Diet:Treatment
aov.2<-aov (model.2, data=dades)
summary(aov.2) # this shows that interaction had to be considered
```

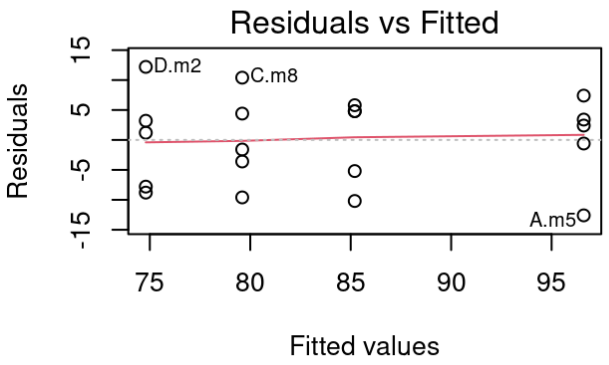
```
               Df Sum Sq Mean Sq F value  Pr(>F)
Diet            1  0.882   0.882   3.039  0.1005
Treatment       1  1.458   1.458   5.023  0.0395 *
Diet:Treatment  1  8.978   8.978  30.932 4.3e-05 ***
Residuals      16  4.644   0.290
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

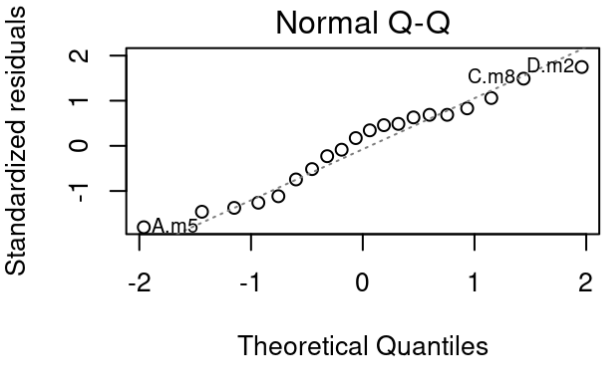
```
with (dades, interaction.plot(Diet, Treatment, WBC))
```

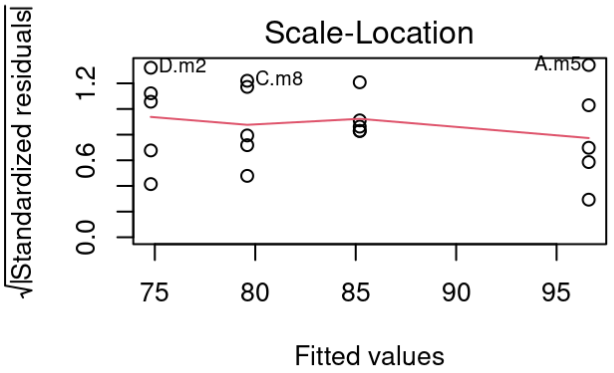


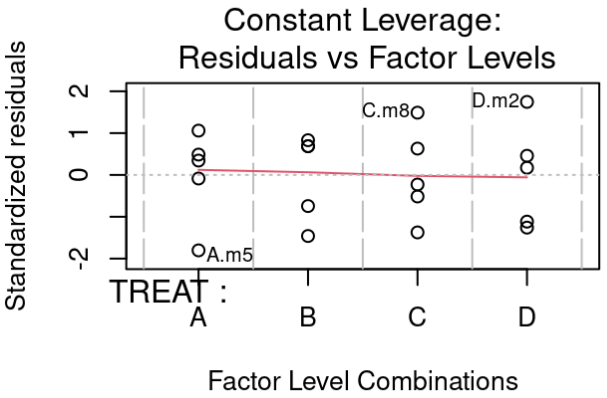
Check ANOVA model assumptions

```
opt<- par(mfrow=c(2,2)); plot(aov1); par(opt)
```


Residuals vs Fitted
Residuals
Fitted values
D.m2
C.m8
A.m5


Normal Q-Q
Standardized residuals
Theoretical Quantiles
C.m8
D.m2
A.m5


Scale-Location
√|Standardized residuals|
Fitted values
D.m2
C.m8
A.m5


Constant Leverage:
Residuals vs Factor Levels
Standardized residuals
Factor Level Combinations
TREAT :
C.m8
D.m2
A.m5