

Analysis of variance ANOVA

André Meichtry

Contents

| | |
|--|----------|
| Univariate one-factorial ANOVA | 1 |
| Statistical model | 1 |
| Example Data | 1 |
| ANOVA as model comparison | 3 |
| Calculations “by hand” | 3 |
| Sum of squares | 3 |
| Degrees of freedom | 3 |
| F -test | 3 |
| Estimates | 4 |
| Contrasts | 4 |
| Check assumptions | 5 |
| Classical version | 5 |
| Two-factorial ANOVA | 6 |
| Statistical model | 6 |
| Simulate some data | 6 |
| Two-factorial model | 7 |
| Sequential versus marginal effects | 8 |

```
library(psych)
library(emmeans)
```

Univariate one-factorial ANOVA

Statistical model

1. Means parameterization:

- $Y_{ij} = \mu_i + \epsilon_{ij}$, $i = 1, \dots, I$; $j = 1, \dots, n_i$; $\epsilon_{ij} \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$ (or with ϵ_i i.i.d. and “large” sample size)
- $Y_{ij} \stackrel{i.i.d.}{\sim} N(\mu_i, \sigma^2)$ (equivalent)

2. Effects parameterization:

- $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$
- $Y_{ij} \stackrel{i.i.d.}{\sim} N(\mu + \alpha_i, \sigma^2)$ (equivalent)

Example Data

We reproduce the example in AMT, page 99.

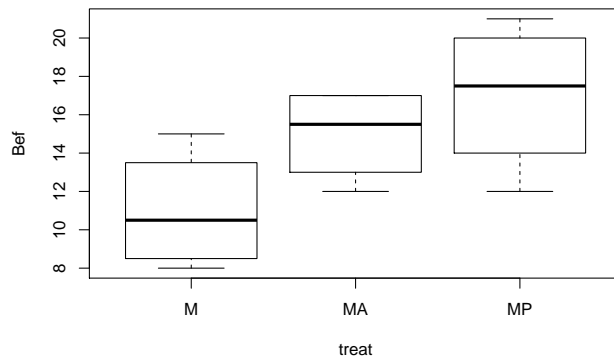
```
Bef <- c(20, 12, 18, 14, 16, 21, 17, 13, 18, 21, 13, 12, 15, 17, 16, 17, 9, 10, 15, 8, 8, 11, 13, 14)
treat<-as.factor(c(rep("MP",10),rep("MA",6),rep("M",8)))
data<-data.frame(Bef,treat)
headTail(data)
```

```
      Bef treat
1      20    MP
2      12    MP
3      18    MP
4      14    MP
... .. <NA>
21      8     M
22     11     M
23     13     M
24     14     M
```

```
str(data)
```

```
'data.frame':  24 obs. of  2 variables:
 $ Bef  : num  20 12 18 14 16 21 17 13 18 21 ...
 $ treat: Factor w/ 3 levels "M","MA","MP": 3 3 3 3 3 3 3 3 3 3 ...
```

```
boxplot(Bef~treat,data=data)
```



```
describeBy(data,group=data$treat,mat=FALSE)
```

Descriptive statistics by group

group: M

| | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
|--------|------|---|------|------|--------|---------|------|-----|-----|-------|------|----------|------|
| Bef | 1 | 8 | 11 | 2.73 | 10.5 | 11 | 3.71 | 8 | 15 | 7 | 0.22 | -1.79 | 0.96 |
| treat* | 2 | 8 | 1 | 0.00 | 1.0 | 1 | 0.00 | 1 | 1 | 0 | NaN | NaN | 0.00 |

group: MA

| | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
|--------|------|---|------|-----|--------|---------|------|-----|-----|-------|-------|----------|------|
| Bef | 1 | 6 | 15 | 2.1 | 15.5 | 15 | 2.22 | 12 | 17 | 5 | -0.33 | -1.88 | 0.86 |
| treat* | 2 | 6 | 2 | 0.0 | 2.0 | 2 | 0.00 | 2 | 2 | 0 | NaN | NaN | 0.00 |

group: MP

| | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
|--------|------|----|------|------|--------|---------|------|-----|-----|-------|-------|----------|------|
| Bef | 1 | 10 | 17 | 3.23 | 17.5 | 17.12 | 4.45 | 12 | 21 | 9 | -0.18 | -1.57 | 1.02 |
| treat* | 2 | 10 | 3 | 0.00 | 3.0 | 3.00 | 0.00 | 3 | 3 | 0 | NaN | NaN | 0.00 |

ANOVA as model comparison

- We will look at ANOVAS from the perspective of **model comparison**.
- Unrestricted model: μ_1, μ_2, μ_3 have to be estimated:

```
mod <- lm(Bef~treat,data)
```

- Restricted model: $\mu_1 = \mu_2 = \mu_3$, or $\alpha_i = 0$. There remains only one parameter, the overall intercept μ :
 $Y_i = \mu + \epsilon_i$

```
mod0 <- lm(Bef~1,data)
```

- Test $\alpha_i = 0, i = 1, 2, 3$

```
anova(mod0,mod)
```

Analysis of Variance Table

Model 1: Bef ~ 1

Model 2: Bef ~ treat

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|--------|-----|----|-----------|------|---------|
| 1 | 23 | 330 | | | | |
| 2 | 21 | 168 | 2 | 162 | 10.1 | 0.00083 |

Calculations “by hand”

Sum of squares

```
(RSS <-sum(residuals(mod)^2))
```

```
[1] 168
```

```
(SS <-sum(residuals(mod0)^2))
```

```
[1] 330
```

```
((SS-RSS)/SS) ##known as R^2
```

```
[1] 0.49091
```

Degrees of freedom

```
(df2<-mod$df.residual) ##numerator df
```

```
[1] 21
```

```
(df1<-mod0$df.residual-mod$df.residual) ##denominator df
```

```
[1] 2
```

F-test

The F -statistic is the **amount of available fit that is actually achieved**, that is

$$F = \frac{(SS - RSS)/df_1}{RSS/df_2} = \frac{MS_{explained}}{MS_{error}}$$

```
F <- (SS-RSS)/(df1)/(RSS/(df2))
p <- 1-pf(F,df1=df1,df2=df2)
sigma.mod <- sqrt(RSS/df2)
print(data.frame(SS,RSS,ESS=SS-RSS,F,p,sigma.mod),row.names=FALSE)
```

```
SS RSS ESS      F      p sigma.mod
330 168 162 10.125 0.00083436 2.8284
```

Estimates

```
summary(mod)
```

Call:

```
lm(formula = Bef ~ treat, data = data)
```

Residuals:

```
Min      1Q  Median      3Q      Max
-5.00  -2.25   0.00   2.00   4.00
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    11.00      1.00    11.00 3.6e-10
treatMA         4.00      1.53     2.62 0.01605
treatMP         6.00      1.34     4.47 0.00021
```

Residual standard error: 2.83 on 21 degrees of freedom

Multiple R-squared: 0.491, Adjusted R-squared: 0.442

F-statistic: 10.1 on 2 and 21 DF, p-value: 0.000834

```
confint(mod)
```

```
              2.5 % 97.5 %
(Intercept) 8.92039 13.0796
treatMA      0.82334 7.1767
treatMP      3.20991 8.7901
```

```
#summary(mod0) ## The only-intercept model, uncomment if you want to look at it.
```

Contrasts

We want to estimate $\mu_1 - \mu_2$, etc.

```
emmeans(mod,specs=pairwise~treat,infer=TRUE) ##Estimated marginal means
```

```
$emmeans
```

```
treat emmean SE df lower.CL upper.CL t.ratio p.value
M      11 1.000 21 8.92 13.1 11.000 <.0001
MA     15 1.155 21 12.60 17.4 12.990 <.0001
MP     17 0.894 21 15.14 18.9 19.007 <.0001
```

Confidence level used: 0.95

```
$contrasts
```

| contrast | estimate | SE | df | lower.CL | upper.CL | t.ratio | p.value |
|----------|----------|------|----|----------|----------|---------|---------|
| M - MA | -4 | 1.53 | 21 | -7.85 | -0.15 | -2.619 | 0.0408 |
| M - MP | -6 | 1.34 | 21 | -9.38 | -2.62 | -4.472 | 0.0006 |
| MA - MP | -2 | 1.46 | 21 | -5.68 | 1.68 | -1.369 | 0.3744 |

Confidence level used: 0.95

Conf-level adjustment: tukey method for comparing a family of 3 estimates

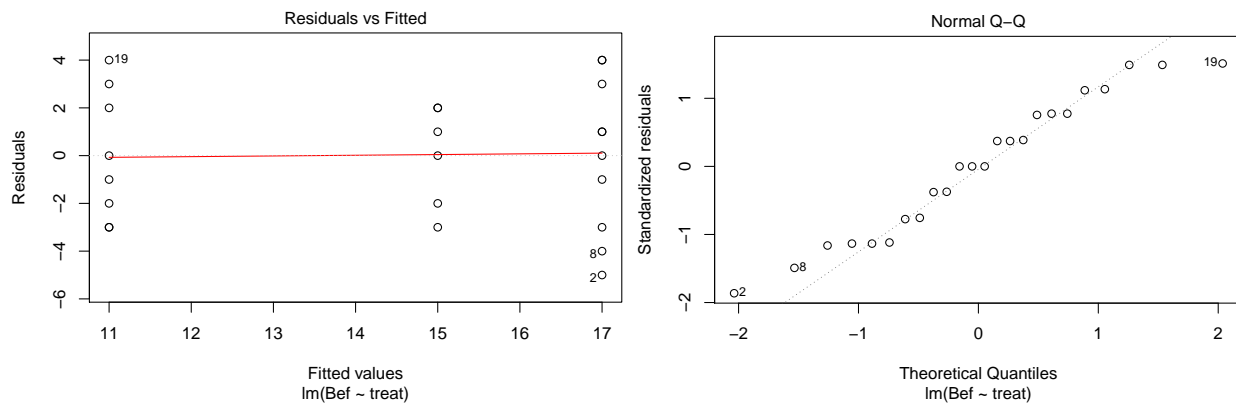
P value adjustment: tukey method for comparing a family of 3 estimates

Check assumptions

Our data were simulated from an ANOVA-modell, so the assumptions are met.

In the analysis stage, however, we always have to check the assumption of **homogeneity of variance** (and normality, but less important).

```
plot(mod,which=c(1,2))
```



Classical version

aov() function in R:

```
modc <-aov(Bef~treat,data)
summary(modc)
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|----|--------|---------|---------|---------|
| treat | 2 | 162 | 81 | 10.1 | 0.00083 |
| Residuals | 21 | 168 | 8 | | |

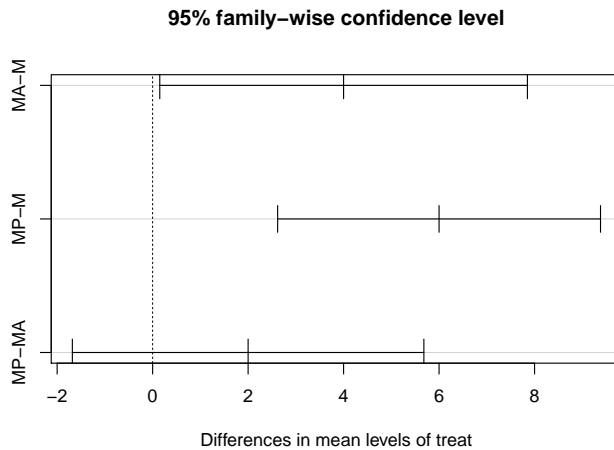
```
TukeyHSD(modc)
```

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

```
Fit: aov(formula = Bef ~ treat, data = data)
```

| | diff | lwr | upr | p adj |
|-------|------|----------|--------|---------|
| MA-M | 4 | 0.14977 | 7.8502 | 0.04081 |
| MP-M | 6 | 2.61830 | 9.3817 | 0.00059 |
| MP-MA | 2 | -1.68153 | 5.6815 | 0.37443 |

```
plot(TukeyHSD(modc))
```



Two-factorial ANOVA

Statistical model

2. Effects parameterization:

- $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, $i = 1, \dots, I$, $j = 1, \dots, J$, $k = 1, \dots, n_{ij}$, $\epsilon_{ijk} \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$.
- $Y_{ijk} \stackrel{i.i.d.}{\sim} N(\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}, \sigma^2)$ (equivalent)

3. Means parameterization:

- $Y_{ijk} = \mu_{ij} + \epsilon_{ijk}$
- $Y_{ijk} \stackrel{i.i.d.}{\sim} N(\mu_{ij}, \sigma^2)$ (equivalent)

Simulate some data

We simulate some cross-sectional data. You do not need to understand the code.

```
nage <- 3
ntherapy <- 2
nsample <- 100
n <- nage * nsample * ntherapy
age <- gl(n = nage, k = nsample, length = n, labels=c("child", "young", "old"))
therapy <- gl(n = ntherapy, k = nsample, length = n, labels=c("Ctrl", "Trt"))
mu <- 40
alpha <- c(1, 1)
beta <- c(1)
gamma <- c(-3, 3)
parameter <- c(mu, alpha, beta, gamma)
sigma <- 12
set.seed(9)
eps <- rnorm(n, 0, sigma)
X <- as.matrix(model.matrix(~ age*therapy))
response <- as.numeric(as.matrix(X) %*% as.matrix(parameter) + eps)
d.cross <- data.frame(response, age, therapy)
```

```
headTail(d.cross)
```

```

      response  age therapy
1       30.8 child   Ctrl
2       30.2 child   Ctrl
3       38.3 child   Ctrl
4       36.67 child   Ctrl
...      ...  <NA>   <NA>
597     38.57  old    Trt
598     75.22  old    Trt
599     27.52  old    Trt
600     27.97  old    Trt

```

Two-factorial model

We begin with a cross-sectional analysis, a two-factorial *balanced* design:

```
with(d.cross,xtabs(~age+therapy))
```

```

      therapy
age    Ctrl Trt
child  100 100
young  100 100
old    100 100

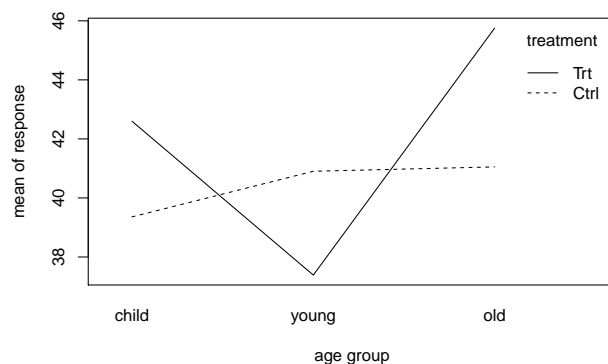
```

```
aggregate(response~therapy+age,data=d.cross,summary)
```

| | therapy | age | response.Min. | response.1st Qu. | response.Median | response.Mean |
|---|---------|-------|---------------|------------------|-----------------|---------------|
| 1 | Ctrl | child | 8.5875 | 31.0511 | 37.9505 | 39.3578 |
| 2 | Trt | child | 16.9479 | 34.6612 | 42.0354 | 42.6000 |
| 3 | Ctrl | young | 4.5071 | 34.4806 | 41.0679 | 40.9037 |
| 4 | Trt | young | 6.1445 | 28.7448 | 39.1686 | 37.3866 |
| 5 | Ctrl | old | 11.3583 | 34.9932 | 40.3621 | 41.0520 |
| 6 | Trt | old | 15.1560 | 37.4397 | 47.1021 | 45.7549 |

| | response.3rd Qu. | response.Max. |
|---|------------------|---------------|
| 1 | 45.0638 | 72.1839 |
| 2 | 50.3490 | 74.1503 |
| 3 | 46.9820 | 70.0471 |
| 4 | 46.3560 | 60.6683 |
| 5 | 49.4753 | 65.1788 |
| 6 | 52.9712 | 75.2195 |

```
with(d.cross,interaction.plot(x.factor=age,trace.factor=therapy,response=response,trace.label="treatment"))
```



```
model2f <-lm(response~age*therapy,data=d.cross)
summary(model2f)
```

Call:

```
lm(formula = response ~ age * therapy, data = d.cross)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|--------|-------|--------|------|-------|
| -36.40 | -7.84 | 0.03 | 7.47 | 32.83 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|---------------------|----------|------------|---------|----------|
| (Intercept) | 39.36 | 1.17 | 33.52 | <2e-16 |
| ageyoung | 1.55 | 1.66 | 0.93 | 0.3522 |
| ageold | 1.69 | 1.66 | 1.02 | 0.3080 |
| therapyTrt | 3.24 | 1.66 | 1.95 | 0.0513 |
| ageyoung:therapyTrt | -6.76 | 2.35 | -2.88 | 0.0041 |
| ageold:therapyTrt | 1.46 | 2.35 | 0.62 | 0.5342 |

Residual standard error: 11.7 on 594 degrees of freedom

Multiple R-squared: 0.0474, Adjusted R-squared: 0.0394

F-statistic: 5.91 on 5 and 594 DF, p-value: 2.4e-05

Compare to true values

```
parameter
```

```
[1] 40 1 1 1 -3 3
```

```
anova(model2f)
```

Analysis of Variance Table

Response: response

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-------------|-----|--------|---------|---------|--------|
| age | 2 | 1825 | 912 | 6.62 | 0.0014 |
| therapy | 1 | 327 | 327 | 2.37 | 0.1242 |
| age:therapy | 2 | 1923 | 962 | 6.98 | 0.0010 |
| Residuals | 594 | 81878 | 138 | | |

```
summary(aov(response~age*therapy,data=d.cross)) ## equivalent!
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-------------|-----|--------|---------|---------|--------|
| age | 2 | 1825 | 912 | 6.62 | 0.0014 |
| therapy | 1 | 327 | 327 | 2.37 | 0.1242 |
| age:therapy | 2 | 1923 | 962 | 6.98 | 0.0010 |
| Residuals | 594 | 81878 | 138 | | |

Sequential versus marginal effects

Take care: F -tests in R (`anova()` and `aov()`) are **sequential** (so-called Type I sum of squares), the order the terms enter the model does matter! The t -tests in `lm()` on the contrary are marginal (impact of the variables, given the presence of all the other variables in the model). This is important when the design is unbalanced. (This is not the case in our example, the results do not differ). Let us change the order.


```
summary(aov(response~therapy*age,data=d.cross)) ## equivalent!
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-------------|-----|--------|---------|---------|--------|
| therapy | 1 | 327 | 327 | 2.37 | 0.1242 |
| age | 2 | 1825 | 912 | 6.62 | 0.0014 |
| therapy:age | 2 | 1923 | 962 | 6.98 | 0.0010 |
| Residuals | 594 | 81878 | 138 | | |

If one needs so-called Type III sum of squares (marginal effects), you have to use the `Anova()` function of the package **car**.

```
library(car)
Anova(model2f,type=3)
```

Anova Table (Type III tests)

Response: response

| | Sum Sq | Df | F value | Pr(>F) |
|-------------|--------|-----|---------|--------|
| (Intercept) | 154904 | 1 | 1123.79 | <2e-16 |
| age | 176 | 2 | 0.64 | 0.528 |
| therapy | 526 | 1 | 3.81 | 0.051 |
| age:therapy | 1923 | 2 | 6.98 | 0.001 |
| Residuals | 81878 | 594 | | |