vliruos

SHARING MINDS, CHANGING LIVES

This course was developed as a part of the VLIR-UOS Cross-Cutting project s:

- Statistics: 2011-2016, 2017.
- Statistics: 2017.
- Statistics for development : 2018-2020.

The >eR-Biostat initiative

Making R based education materials in statistics accessible for all

An introduction to R: Short Version (2017)

Part 4: statistical modeling 2

Developed by

Dan Lin (Hasselt University) and Ziv Shkedy (Hasselt University)

LAST UPDATE: 15/10/2017

Visit us on Facebook ER-BioStat

GitHub https://github.com/eR-Biostat

Email: erbiostat@gmail.com

twitter @erbiostat

Overview

1. Two-way ANOVA.
2. More about two-way ANOVA.
3. More about linear regression.

Statistical modeling : Two-way ANOVA

Model formulation

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

μ	Overall mean
α_i	Main effect of factor A
β_j	Main effect of factor B
$\alpha\beta_{ij}$	Interaction effect
ε_{ijk}	Random error

Example 1: Reading the data

```
> spwh3<-read.table('c:\\projects\\wseda\\spwh3.txt',
  header=FALSE,na.strings="NA", dec=".")
> names(spwh3)<-c("id","y","x1","gender")
```

Example 1: The data

```
> print(spwh3)
   id         y x1 gender
1   1 10.111368  1      0
2   2  9.948930  1      0
3   3 10.322560  1      0
.   .   .        .      .
.   .   .        .      .
59 59 30.030490  3      1
60 60 29.541542  3      1
>
```

Both x1 and gender are numerical objects !!!!

For an ANOVA model the independent variables are suppose to be factors.

Example 2: The data

```
    y  f1 f2
1  10  A1 B1
2  11  A1 B1
3  12  A1 B1
4   9  A2 B1
5   7  A2 B1
6   6  A2 B1
7  11  A1 B2
8  13  A1 B2
9  14  A1 B2
10  7  A2 B2
11  5  A2 B2
12  8  A2 B2
```

Two factors: f1 and f2

Three observations per combination.

```
> f1<-c("A1","A1","A1","A2","A2","A2","A1","A1","A1","A2","A2","A2")
> f2<-c("B1","B1","B1","B1","B1","B1","B2","B2","B2","B2","B2","B2")
> y<-c(10,11,12,9,7,6,11,13,14,7,5,8)
> data.frame(y,f1,f2)
```

Which null hypotheses we test ?

$$H_0 : \alpha_1 = \alpha_2$$

No treatment effect of factor A

$$H_0 : \beta_1 = \beta_2$$

No treatment effect of factor B

No interaction effects

$$H_0 : \alpha\beta_{11} = \alpha\beta_{12} = \alpha\beta_{21} = \alpha\beta_{22}$$

Example 1: A model without interaction

```
> fit.1<-aov(y~as.factor(x1)+as.factor(gender))
> anova(fit.1)
Analysis of Variance Table

Response: y
                  Df  Sum Sq Mean Sq F value    Pr(>F)    
as.factor(x1)      2 1034.81  517.40  2244.8 < 2.2e-16 ***
as.factor(gender)  1 1509.98 1509.98  6551.3 < 2.2e-16 ***
Residuals         56   12.91    0.23                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1
```

Example 1: A model with interaction

```
fit.2<-aov(y~as.factor(x1)+as.factor(gender)
           +as.factor(x1)*as.factor(gender))
```

```
> anova(fit.2)
Analysis of Variance Table

Response: y
                                Df  Sum Sq Mean Sq  F value Pr(>F)
as.factor(x1)                    2 1034.81  517.40 2171.959 <2e-16 ***
as.factor(gender)                1 1509.98 1509.98 6338.599 <2e-16 ***
as.factor(x1):as.factor(gender)  2    0.04    0.02    0.091 0.9131
Residuals                       54   12.86    0.24
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Example 1: Testing model 1 versus model 2

```
> anova(fit.1,fit.2)
```

```
Analysis of Variance Table
Model 1: y ~ as.factor(x1) + as.factor(gender)
Model 2: y ~ as.factor(x1) + as.factor(gender) + as.factor(x1) * as.factor(gender)
```

```
   Res.Df    RSS     Df   Sum of Sq       F     Pr(>F)
1      56 12.9073
2      54 12.8639   2     0.0434      0.091   0.9131
```

F-test for the interaction

Example 2: A model without interaction

```
> fit.1<-aov(y~f1+f2)
> anova(fit.1)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
f1         1 70.083  70.083 31.4066 0.0003325 ***
f2         1  0.750   0.750  0.3361 0.5763122
Residuals  9 20.083   2.231
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
```

Example 2: A model with interaction

```
> fit.2<-aov(y~f1+f2+f1*f2)
> anova(fit.2)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
f1         1 70.083  70.083 35.0417 0.0003539 ***
f2         1  0.750   0.750  0.3750 0.5572922    
f1:f2      1  4.083   4.083  2.0417 0.1909016    
Residuals  8 16.000   2.000                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example 2: Testing model 1 versus model 2

```
> anova(fit.1,fit.2)
```

```
Analysis of Variance Table

Model 1: y ~ f1 + f2
Model 2: y ~ f1 + f2 + f1 * f2
```

```
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      9 20.083
2      8 16.000  1    4.0833 2.0417 0.1909
```

F-test for the interaction

Example 2: means by factor level

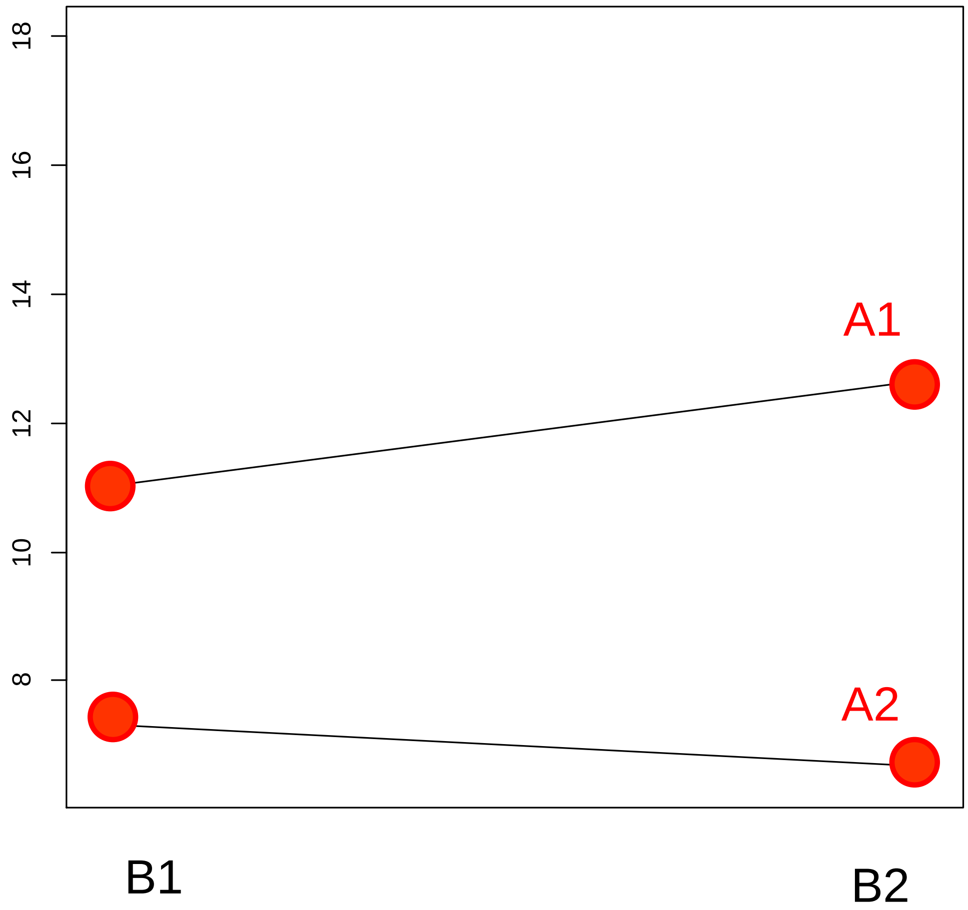
```
> tapply(y,f1,mean)
      A1       A2
11.83333 12.00000
> tapply(y,f2,mean)
       B1        B2
 9.166667 14.666667
> ind<-list(f1,f2)
> ind
[[1]]
 [1] "A1" "A1" "A1" "A2" "A2" "A2" "A1" "A1" "A1" "A2" "A2" "A2"

[[2]]
 [1] "B1" "B1" "B1" "B1" "B1" "B1" "B2" "B2" "B2" "B2" "B2" "B2"

> m<-tapply(y,ind,mean)
> m
          B1       B2
A1 11.000000 12.66667
A2  7.333333 16.66667
```

Factor 1

Factor 2

Cell means

Interaction plot: Example 2



Cell means

	B1	B2
A1	11.000000	12.666667
A2	7.333333	6.666667

Example 3: The data

```
   y f1 f2
1  10 A1 B1
2  11 A1 B1
3  12 A1 B1
4   9 A2 B1
5   7 A2 B1
6   6 A2 B1
7  11 A1 B2
8  13 A1 B2
9  14 A1 B2
10 17 A2 B2
11 15 A2 B2
12 18 A2 B2
```

Two factors: f1 and f2

Three observations per combination.

```
> f1<-c("A1","A1","A1","A2","A2","A2","A1","A1","A1","A2","A2","A2")
> f2<-c("B1","B1","B1","B1","B1","B1","B2","B2","B2","B2","B2","B2")
> y<-c(10,11,12,9,7,6,11,13,14,17,15,18)
> data.frame(y,f1,f2)
```

Example 3: A model with interaction

```
> fit.2<-aov(y~f1+f2+f1*f2)
> anova(fit.2)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
f1         1  0.083   0.083  0.0417 0.8433536
f2         1 90.750  90.750 45.3750 0.0001471 ***
f1:f2      1 44.083  44.083 22.0417 0.0015517 **
Residuals  8 16.000   2.000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example 3: means by factor level

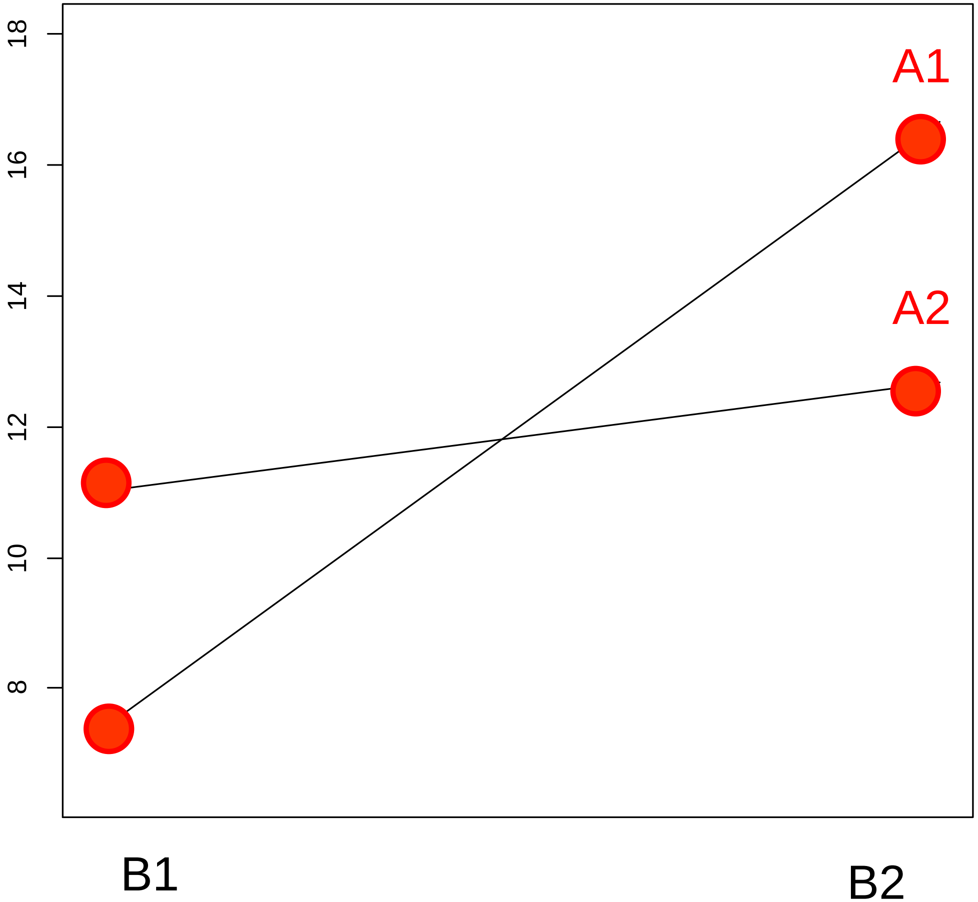
```
> tapply(y,f1,mean)
      A1       A2
11.83333 12.00000
> tapply(y,f2,mean)
       B1        B2
 9.166667 14.666667
> ind<-list(f1,f2)
> ind
[[1]]
 [1] "A1" "A1" "A1" "A2" "A2" "A2" "A1" "A1" "A1" "A2" "A2" "A2"

[[2]]
 [1] "B1" "B1" "B1" "B1" "B1" "B1" "B2" "B2" "B2" "B2" "B2" "B2"

> m<-tapply(y,ind,mean)
> m
          B1       B2
A1 11.000000 12.66667
A2  7.333333 16.66667
```

Factor 1

Factor 2

Cell means

Interaction plot: Example 3



Cell means

	B1	B2
A1	11.000000	12.66667
A2	7.333333	16.66667

Statistical modeling :
More about two-way ANOVA

Reading the data

```
> spwh3<-read.table('c:\\projects\\wseda\\spwh3.txt',
header=FALSE,na.strings="NA", dec=".")
> names(spwh3)<-c("id","y","x1","gender")
> attach(spwh3)
```

Two-way ANOVA model

```
> fit.2<-aov(y~as.factor(x1)+as.factor(gender)+as.factor(x1)*as.factor(gender))

> anova(fit.2)

Analysis of Variance Table

Response: y
                                Df  Sum Sq Mean Sq  F value Pr(>F)
as.factor(x1)                    2 1034.81  517.40 2171.959 <2e-16 ***
as.factor(gender)                1 1509.98 1509.98 6338.599 <2e-16 ***
as.factor(x1):as.factor(gender)  2    0.04    0.02    0.091 0.9131
Residuals                       54   12.86    0.24
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Stepwise procedure

```
> slm1 <- step(fit.2)
Start:  AIC=-80.4
y ~ as.factor(x1) + as.factor(gender) + as.factor(x1) * as.factor(gender)

                                   Df Sum of Sq     RSS     AIC
- as.factor(x1):as.factor(gender)  2     0.043  12.907 -84.193
<none>                                           12.864 -80.395

Step:  AIC=-84.19
y ~ as.factor(x1) + as.factor(gender)

                    Df Sum of Sq     RSS     AIC
<none>                             12.91  -84.19
- as.factor(x1)      2   1034.81 1047.72  175.60
- as.factor(gender)  1   1509.98 1522.89  200.04
```

Stepwise procedure

```
> summary(slm1)
                   Df  Sum Sq Mean Sq F value    Pr(>F)    
as.factor(x1)       2 1034.81  517.40  2244.8 < 2.2e-16 ***
as.factor(gender)   1 1509.98 1509.98  6551.3 < 2.2e-16 ***
Residuals          56   12.91    0.23                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
```

Statistical modeling :
More about Linear regression

Reading the data

```
> spwh2<-read.table('c:\\projects\\wseda\\spwh2.txt',
header=FALSE,
+                            ,na.strings="NA", dec=".")
> dim(spwh2)
[1] 100   5
>
> names(spwh2)<-c("id","y","x1","x2","x3")
> attach(spwh2)

        The following object(s) are masked from spwh2 (
position 3 ) :

         id x1 x2 x3 y
```

Fitting two models

```
> fit.1<-lm(y~x1+x2)
> anova(fit.1)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
x1         1  164.2   164.2   27.152 1.059e-06 ***
x2         1 7409.7  7409.7 1224.980 < 2.2e-16 ***
Residuals 97  586.7     6.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
     ‘.’ 0.1 ‘ ’ 1
```

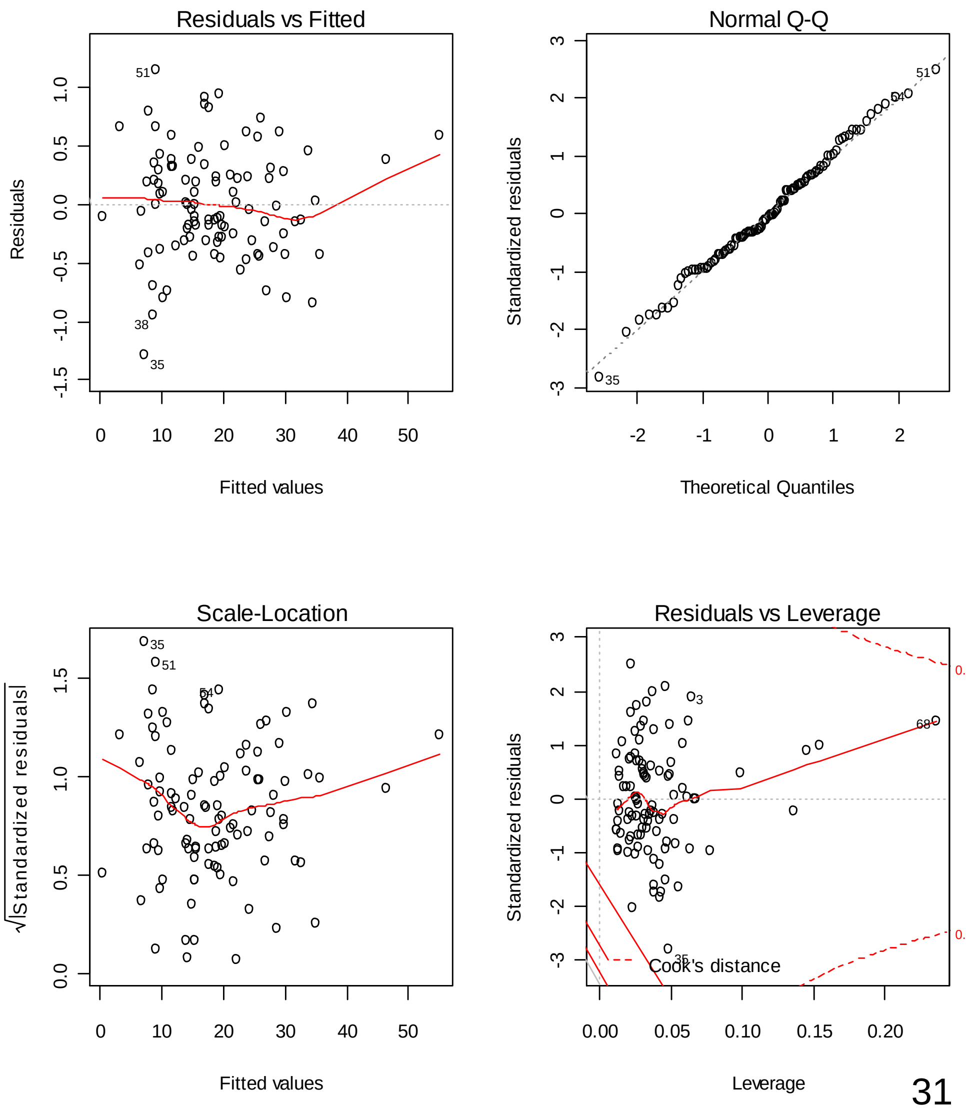
```
> fit.2<-lm(y~x1+x2+x3)
> anova(fit.2)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
x1         1  164.2   164.2   758.98 < 2.2e-16 ***
x2         1 7409.7  7409.7 34241.81 < 2.2e-16 ***
x3         1  566.0   566.0  2615.44 < 2.2e-16 ***
Residuals 96   20.8     0.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
     0.1 ‘ ’ 1
```

Testing model 1 versus model 2

```
> anova(fit.1,fit.2)
Analysis of Variance Table

Model 1: y ~ x1 + x2
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     97 586.74
2     96  20.77  1    565.97 2615.4 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1
```

```
> par(mfrow=c(2,2))
> plot(fit.2)
```



Single terms deletions

```
> drop1(fit.2, test="F")
Single term deletions

Model:
y ~ x1 + x2 + x3
       Df Sum of Sq    RSS    AIC  F value     Pr(F)
<none>               20.8 -149.1
x1      1      76.6   97.4    3.4   354.21 < 2.2e-16 ***
x2      1    7865.3 7886.1  442.8 36347.01 < 2.2e-16 ***
x3      1     566.0  586.7  182.9  2615.44 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1
```

AIC and likelihood

```
> AIC(fit.2)
[1] 136.6403
> logLik(fit.2)
'log Lik.' -63.32017 (df=5)
```