Lab 6

!!Please read C10 and C11.

One-way ANOVA, two-way ANOVA

-compare 3 or more population means at once.

One factor

Let's imagine the following experiment: we have four age groups (denoted 1,2,3,4) and 5 persons in each group. We ask each person to visit a web page (the same for everyone) and we record the number of seconds spent on the web page by each participant. We are interested in whether the age influences the amount of time spent on the webpage.

```
#generate data – number of seconds
g1<- round(rnorm(5,178,1.5))
g2<- round(rnorm(5,190,2))
g3<- round(rnorm(5,183,3))
g4<- round(rnorm(5,180,2.4))

d<-list(g1,g2,g3,g4)

bartlett.test(d)   #test for homogeneity of data -- see C10 slides 7,8
```

How do you interpret the outcomes of this test? (in terms of the p-value)

```
time<-c(g1,g2,g3,g4)
agegroup<-rep(c(1,2,3,4),c(5,5,5,5))

factoragegroup<-factor(agegroup)  # agegroup has to be factor – required
                                  # by aov function below
#attention: you get a different result (different dfs) if you use agegroup instead of
# factor(agegroup) – that is because the models based on numerics (quantitative
# data) or factors (qualitative data) differ
# https://stackoverflow.com/questions/21226069/when-are-factors-necessary-appropriate-in-r

mydata<-data.frame(time,factoragegroup)

levels(mydata$factoragegroup)  #check the levels of the factor


a<-aov(formula=time~factoragegroup,data=mydata)
#left_handside ~ right_handside in a formula
#left_handside and right_handside depend on the function that has
# these parameters --- in our case aov
```

```
summary(a)

                   Df Sum Sq Mean Sq F value   Pr(>F)    
factoragegroup  3    447  148.98   20.91 8.79e-06 ***
Residuals      16    114    7.13                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Obs. Your result will be different as the data are randomly generated.

In the ANOVA table returned by summary(a), we see:
-The row “factoragegroup” corresponding to SS_α (C10 slide 14)
df=3 (that is r-1; r=4 is the no. of levels of the factor)
-The row “Residuals” corresponding to SS_r
df=16 (that is n-r; n=20 the sum of all sample sizes).

-Pr(>F) is the p-value of the F-test – we reject the null hypothesis – the factor “agegroup” has a significant influence over the amount of time spent on the webpage.

!! Obs. The test does not tell us which groups are different.

Generate now g2<- round(rnorm(5,181,2)) (181 instead of 190) and run the ANOVA again. What is the result now, according to the ANOVA table?

Two factors

Let’s imagine now the following experiment: we have two drugs (factors) for blood pressure, that are administered together. Each drug is given in two doses (levels). We are interested in whether the treatments (all the 4 combinations) influence the blood pressure of different patients. We assume that we have 5 patients in each group (balanced experiments).

```
#generate data – blood pressure
g1<- round(rnorm(5,10,2))
g2<- round(rnorm(5,12,2))
g3<- round(rnorm(5,13,2))
g4<- round(rnorm(5,15,2))

d<-list(g1,g2,g3,g4)

bartlett.test(d)

bloodpressure<-c(g1,g2,g3,g4)
treat1<-rep(c(1,2),c(10,10))
treat2<-rep(rep(c(1,2),c(5,5)),2)

factortreat1<-factor(treat1)
factortreat2<-factor(treat2)
```

```
mydata<-data.frame(bloodpressure,factortreat1, factortreat2)

levels(mydata$factortreat1)  #check the levels
levels(mydata$factortreat2)

a<-aov(formula=bloodpressure~ factortreat1*factortreat2,data=mydata)
#or equivalent bloodpressure~ factortreat1+factortreat2+ factortreat1:factortreat2
#we use that when we think that the 2 factors may interact


summary(a)
```

```
                          Df Sum Sq Mean Sq F value Pr(>F)
factortreat1               1  42.05   42.05   8.205 0.0112 *
factortreat2               1  14.45   14.45   2.820 0.1125
factortreat1:factortreat2  1  14.45   14.45   2.820 0.1125
Residuals                 16  82.00    5.12
```

In the ANOVA table returned by summary(a), we see:

-The row “factortreat1” corresponding to SS_α (C10 slide 22) and H_α
df=1 (that is r-1; r=2 no of levels of the 1st factor)

-The row “factortreat2” corresponding to SS_β and H_β
df=1 (that is s-1; s=2 no of levels of the 2nd factor)

-The row “factortreat1:factortreat2” corresponding to SS_γ and H_γ
df=1 (that is (r-1)*(s-1)))

-The row “Residuals” corresponding to SS_r
df=16 (that is r*s*(n-1); n=5 no. of observations in each group).

-Pr(>F) are the p-values of the F-tests – we reject H_α , but we fail to reject H_β and H_γ – the factor “treat1” has a significant influence (effect) on the blood pressure; the factor “treat2” does not have a significant influence on the blood pressure; and there are no significant interaction between the two factors.

Linear regression

Example 1 – k=1 – regression line

```
longley
X <- longley[, "Employed"]     # number of people employed
Y <- longley[,"Population"]    # population ≥ 14 years of age

#the sample size n=16

model1<-lm(X~Y)       #X is the “effect”; Y is the “cause”
#C11

model1
```

```
          Call:
          lm(formula = X ~ Y)

          Coefficients:
          (Intercept)            Y
               8.3807       0.4849
```

#the estimation of β_0 is 8.3807
#the estimation of β_1 is 0.4849

#the regression line: $x=8.3807+0.4849*y_1$

```
#residuals(model1)

summary(model1)
```

```
          Call:
          lm(formula = X ~ Y)

          Residuals:
              Min       1Q  Median       3Q      Max
          -1.4362 -0.9740  0.2021  0.5531  1.9048

          Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
          (Intercept)  8.3807     4.4224   1.895   0.0789 .
          Y            0.4849     0.0376  12.896 3.69e-09 ***
          ---
          Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#p-value=3.69e-09 ==> we reject the hypothesis H_0:{ β_1=0}
# see T-test (C11 slide 17)

```
          Residual standard error: 1.013 on 14 degrees of freedom
```

```
Multiple R-squared:  0.9224,     Adjusted R-squared:  0.9168
F-statistic: 166.3 on 1 and 14 DF,  p-value: 3.693e-09
```

```
#p-value=3.69e-09 ==> we reject the hypothesis H0:{ β1=0}
#see F-test (C11 slides 15,16)

#Multiple R-squared:  0.9224 ==> the model fits very well the data (C11 slide 14)



plot(X~Y, ylim = c(5,80))
abline(model1)

#we perform ANOVA, under the assumption of normality
anova(model1)
```

```
Analysis of Variance Table

Response: X
          Df  Sum Sq Mean Sq F value    Pr(>F)
Y          1 170.643 170.643   166.3 3.693e-09 ***
Residuals 14  14.366   1.026
```

-The row “Y” corresponding to $SS_{regression}$ -- df=1 (C11 slide 13)
-The row “Residuals” corresponding to $SS_{residual}$ -- df=14 (=n-2).

We reject the hypothesis H:{ β_1=0}, hence the linear model that we assumed lm(X ~ Y) is significant.

```
#the sample correlation coefficient (C11 slide 20)
cor.test(X,Y)
```

```
        Pearson's product-moment correlation

data:  X and Y
t = 12.896, df = 14, p-value = 3.693e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8869236 0.9864676
sample estimates:
      cor
0.9603906
```

```
# r=0.9603906 – strong correlation between cause and effect -- X increases as Y
# increases
```

Example 2 – k=2 – regression plane

```
longley
x<-longley[,"Employed"]        # number of people employed
y1<-longley[,"GNP"]            # Gross National Product
y2<-longley[,"Year"]           # the year

#the sample size n=16

model2=lm(x~y1+y2)              # x is the effect, $y_1$ and $y_2$ are the causes
model2
```

```
Call:
lm(formula = x ~ y1 + y2)

Coefficients:
(Intercept)           y1           y2
 1198.70811      0.06299     -0.59238
```

```
#the estimation of $\beta_0$ is 1198.70811
#the estimation of $\beta_1$ is 0.06299
#the estimation of $\beta_2$ is -0.59238

#the regression plane: x=1198.70811+0.06299*y1 -0.59238*y2


summary(model2)
```

```
Call:
lm(formula = x ~ y1 + y2)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.8553 -0.3224 -0.1092  0.2369  1.4455 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  1198.70811  664.52142   1.804  0.09446 . 
y1              0.06299    0.01644   3.831  0.00208 **
y2             -0.59238    0.34324  -1.726  0.10805   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.6146 on 13 degrees of freedom
Multiple R-squared:  0.9735,     Adjusted R-squared:  0.9694 
F-statistic: 238.4 on 2 and 13 DF,  p-value: 5.699e-11
```

The overall F-test of the model H_0:{ β_1= β_2=0}
F statistic has 2 and 13 df (that is k-1,n-k-1)
p-value: 5.699e-11 ==> H_0 is rejected

The coefficients are also tested individually H:{ β_1=0}, respectively H:{ β_2=0} with T-tests:
-In the row “y1” above, the p-value is 0.00208 ==> H:{ β_1=0} is rejected
-In the row “y2” above, the p-value is 0.10805 ==> H:{ β_2=0} is accepted

The same conclusions are obtained by using F-tests -- ANOVA for the regression plane (under the assumption of normality).

```
anova(model2)
```

```
Analysis of Variance Table

Response: x
          Df  Sum Sq Mean Sq  F value    Pr(>F)    
y1         1 178.973 178.973 473.7674 1.302e-11 ***
y2         1   1.125   1.125   2.9786     0.108    
Residuals 13   4.911   0.378                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hence, we reject the hypothesis of the regression of x in y_2 and we accept the hypothesis of the regression of x in y_1.
This analysis indicates that the cause y_1 should be retained in the model and the cause y_2 could be dropped.

Therefore, our model will be reduced to:

```
model3=lm(x~y1)
```

Following the steps in the Example 1, analyze now the regression line of model3.