12

Bernat Chiva

12/5/2020

```
library(car)

## Loading required package: carData

library(HH)

## Loading required package: lattice

## Loading required package: grid

## Loading required package: latticeExtra

## Loading required package: multcomp

## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser

## Loading required package: gridExtra

##
## Attaching package: 'HH'

## The following objects are masked from 'package:car':
##
##     logit, vif

library(tables)
library(RcmdrMisc)

## Loading required package: sandwich

library(emmeans)
```

```
##
## Attaching package: 'emmeans'

## The following object is masked from 'package:HH':
##
##     as.glht

library(ggplot2)

##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:latticeExtra':
##
##     layer

library(pander)

library(readr)
PRAC2F <- read_delim("C:/Users/berna/OneDrive/Desktop/UPC/S1/5. Models
Lineals/datasets/PRAC2F.csv",
    ";", escape_double = FALSE, locale = locale(decimal_mark = ","),
    trim_ws = TRUE)

##
## -- Column specification ------------------------------------------------------
------
## cols(
##   F1 = col_double(),
##   F2 = col_double(),
##   V1 = col_double(),
##   V2 = col_double(),
##   V3 = col_double(),
##   V4 = col_double(),
##   V5 = col_double(),
##   V6 = col_double()
## )

PRAC2F$F1 <- as.factor(PRAC2F$F1)
PRAC2F$F2 <- as.factor(PRAC2F$F2)

summ<-summary(PRAC2F)
pander(summ, style='multiline', plan.ascii=F)
```

Table continues below

F1	F2	V1	V2	V3	V4
1:20	1:15	Min. :1.140	Min. :0.870	Min. :1.220	Min. :1.060
2:20	2:15	1st Qu.:1.590	1st Qu.:1.607	1st Qu.:1.657	1st Qu.:1.587
3:20	3:15	Median :1.860	Median :1.915	Median :1.905	Median :1.875
NA	4:15	Mean :1.865	Mean :1.845	Mean :1.878	Mean :1.833

NA	NA	3rd Qu.:2.090	3rd Qu.:2.092	3rd Qu.:2.080	3rd Qu.:2.042
NA	NA	Max. :3.020	Max. :2.600	Max. :2.700	Max. :2.540

V5	V6
Min. :0.960	Min. :0.780
1st Qu.:1.345	1st Qu.:1.330
Median :1.485	Median :1.590
Mean :1.496	Mean :1.576
3rd Qu.:1.675	3rd Qu.:1.740
Max. :2.120	Max. :2.590

We first compute the means table for all the six variables:

```
#means table, descriptive analysis
tab<-tabular((V1+V2+V3+V4+V5+V6)*(F1+1)~mean*(F2+1), PRAC2F)
pander(tab, style='multiline', plan.ascii=F)
```

		mean				
		F2				
\	F1	1	2	3	4	All
V1	*1*	1.748	1.876	1.832	2.480	1.984
	2	1.486	1.688	1.322	1.720	1.554
	3	1.674	1.996	2.490	2.074	2.058
	All	1.636	1.853	1.881	2.091	1.865
V2	*1*	1.544	2.002	2.036	2.116	1.924
	2	1.142	1.676	1.584	1.820	1.555
	3	1.652	2.258	2.072	2.238	2.055
	All	1.446	1.979	1.897	2.058	1.845
V3	*1*	1.652	1.860	1.742	2.244	1.875
	2	1.614	2.008	1.754	2.054	1.857
	3	1.442	1.882	2.304	1.982	1.903
	All	1.569	1.917	1.933	2.093	1.878
V4	*1*	1.488	1.666	1.938	2.052	1.786
	2	1.576	1.854	1.932	2.124	1.871
	3	1.464	1.840	1.952	2.116	1.843

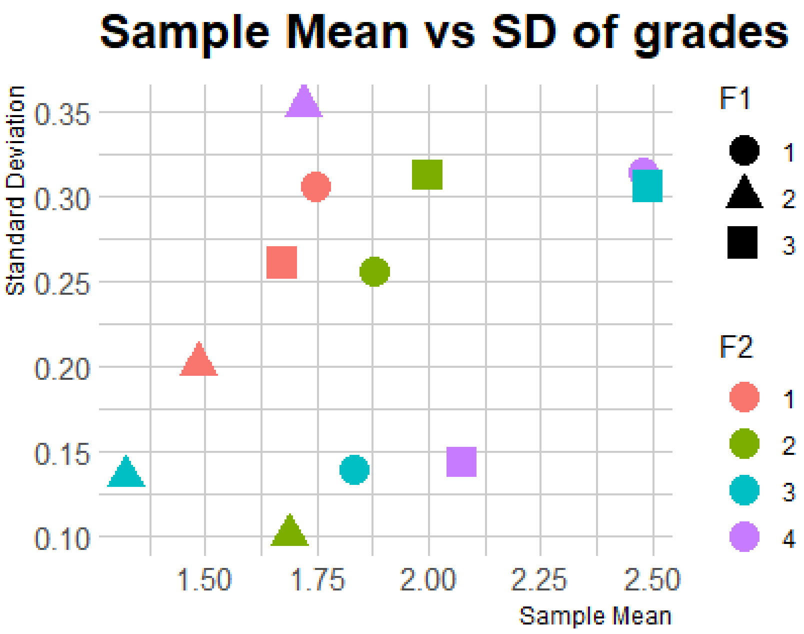
	All	1.509	1.787	1.941	2.097	1.833
V5	*1*	1.670	1.336	1.386	1.692	1.521
	2	1.470	1.690	1.270	1.512	1.486
	3	1.292	1.422	1.972	1.234	1.480
	All	1.477	1.483	1.543	1.479	1.496
V6	*1*	1.486	1.808	1.446	1.640	1.595
	2	1.340	1.544	1.576	1.678	1.534
	3	1.700	1.648	1.418	1.632	1.599
	All	1.509	1.667	1.480	1.650	1.576

1. Interaction models:

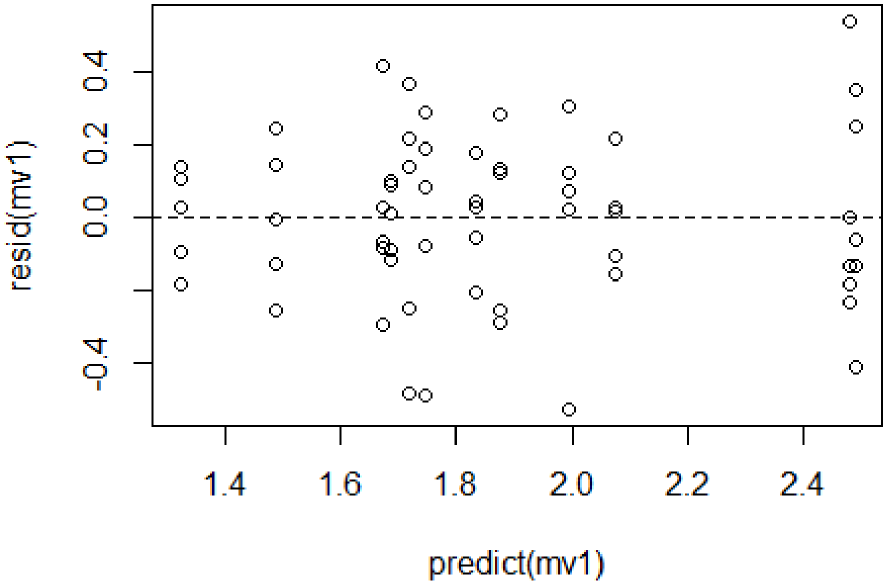
We assume this model: $Y_{ij} = \beta_0 + \beta_1 * F1_i + \beta_2 * F2_j + \beta_3 * F1_i * F2_j$ for our data. We will fir it in R and analyze it for each dependent variable.

1.1. V1

```
#anova test
mv1<-lm(V1~F1*F2, PRAC2F)
anova(mv1)

## Analysis of Variance Table
##
## Response: V1
##           Df Sum Sq Mean Sq F value    Pr(>F)
## F1         2 2.9665 1.48323 23.7530 6.746e-08 ***
## F2         3 1.5610 0.52035  8.3330 0.0001454 ***
## F1:F2      6 2.3321 0.38868  6.2244 6.916e-05 ***
## Residuals 48 2.9973 0.06244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

a<-as.vector(with(PRAC2F, tapply(V1, list(F2=F2, F1=F1),mean)))
b<-as.vector(with(PRAC2F, tapply(V1, list(F2=F2, F1=F1),sd)))

stats<-cbind(a, b)
stats<-data.frame(stats)
stats$F1<-as.factor(c(1,1,1,1,2,2,2,2,3,3,3,3))
stats$F2<-as.factor(c(1,2,3,4,1,2,3,4,1,2,3,4))


ggplot(stats, aes(y=b, x=a, color=F2, shape=F1) )+ xlab("Standard Deviation")
+ylab("Sample Mean")+ ggtitle("Sample Mean vs SD of grades")+
    geom_point(size=5, ) +
    theme_ipsum()
```



```r
plot(predict(mv1),resid(mv1))
abline(h=0,lty=2)
```



a) Explain a real situation (specifying the name of the variables that appear in the experiment and the procedure to obtain the data) that matches with the given file.

The experiment could a model to explain the weight loss by the level of exercise and the group of age of the person.

We should take four groups of ages, 0-21, 22-34, 35-50, +50 and three levels of exercise, intense, regular and casual.

To get a balanced design we should find the right amount of people for each level. The search should be randomized to ensure independency. For instance, we should group all the people we know these two factors from and pick randomly.

• Is there any significant effect? What can be deduced from the means table?

Both factors and the interaction have a significant effect for the first variable. We can deduce the effect and the interaction by checking the variation of the different means in the cells.

• Which is the error variance estimation?

The mean square of the residuals: 0.06244.

• Does the first factor have any effect? How is it detected in the means table and in the X – SX diagram?

F1 has a significant effect. We can detect it in the means table by checking how changing between the three levels of F1, keeping F2 steady, the mean of the variable V1 changes as well.

In the diagram, when checking the effect of F1, we have to observe the different shapes, as each shape represents one level of the first factor. We observe how if we draw a line connecting all the circles, this is not linear, meaning that F1 has an effect on the response variable and the sample means for each level has different value.

• Does the second factor have any effect? How is it detected in the means table and in the X – SX diagram?

F2 has a significant effect. Similar to the last question. We can detect it in the means table by checking how changing between the three levels of F2, keeping F1 steady, the mean of the variable V1 changes as well.

In the diagram, when checking the effect of F2, we have to observe the different colors, as each color represents one level of the second factor. We observe how if we draw a line connecting all the blue shapes, this is not linear, meaning that F2 has an effect on the response variable and the sample means for each level has different value.

- **Does it exists interaction? How is it detected in the means table and in the X – SX diagram?**

Interaction has a significant effect. In the means table we can detect interaction if we see some significant change if we keep F1 at the same level but change F2. For example, if we look at the means table of V1, we would see how the first level of F1, has different values depending on which level F2 is.

In the diagram, to check for interaction, we have to look how the level of each factor gives different means depending on the level of the other factor, this is, how all the circles have a different value, the same for the squares and triangles. We should expect similar values for all the circles if there was no interaction.

- **Which is (are) the best treatment?**

The best treatment is the level 4 for F1 and level 3 for F2, as the expected values is the highest one.

- **Which is (are) the best level of the first factor?**

The best level for F1 is the third, as its mean value upon all the observations is the highest.

- **Which is (are) the best level of the second factor?**

The best level for F2 is the fourth level, as its mean value upon all the observations is the highest.

- **The best treatment, is it the combination of the best levels of each factor? Why?**

No, this is given by the fact that there is interaction and when we combine the variability of the response variable explained both by F1 and F2, the values of each factor depend as well by the level of the other.

- **If the first level of the first factor needs to be used, which treatment is in your opinion the best?**

If we have to use the first level for F1, then the best treatment would be to use the fourth level for F2, as the expected value is the highest upon the means of the first level of F1, checking the first row of the means table.

- **If the second level of the first factor needs to be used, which treatment is in your opinion the best?**

If we have to use the second level for F1, then the best treatment would be again to use the fourth level for F2, as the expected value is the highest upon the means of the first level of F1, checking the second row of the means table.

• If the third level of the first factor needs to be used, which treatment is in your opinion the best?

If we have to use the third level for F1, then the best treatment would be to use the third level for F2, as the expected value is the highest upon the means of the first level of F1, checking the third row of the means table.

• If the first level of the second factor needs to be used, which treatment is in your opinion the best?

If we have to use the first level for F2, then the best treatment would be to use the first level for F1, as the expected value is the highest upon the means of the first level of F2, checking the first column of the means table.

• If the second level of the second factor needs to be used, which treatment is in your opinion the best?

If we have to use the second level for F2, then the best treatment would be to use the third level for F1, as the expected value is the highest upon the means of the first level of F2, checking the second column of the means table.

• If the third level of the second factor needs to be used, which treatment is in your opinion the best?

If we have to use the third level for F2, then the best treatment would be to use the third level for F1, as the expected value is the highest upon the means of the first level of F2, checking the third column of the means table.

• If the fourth level of the second factor needs to be used, which treatment is in your opinion the best?

If we have to use the fourth level for F2, then the best treatment would be to use the first level for F1, as the expected value is the highest upon the means of the first level of F2, checking the fourth column of the means table.

• Do you see any anomalous tendency in the residuals?

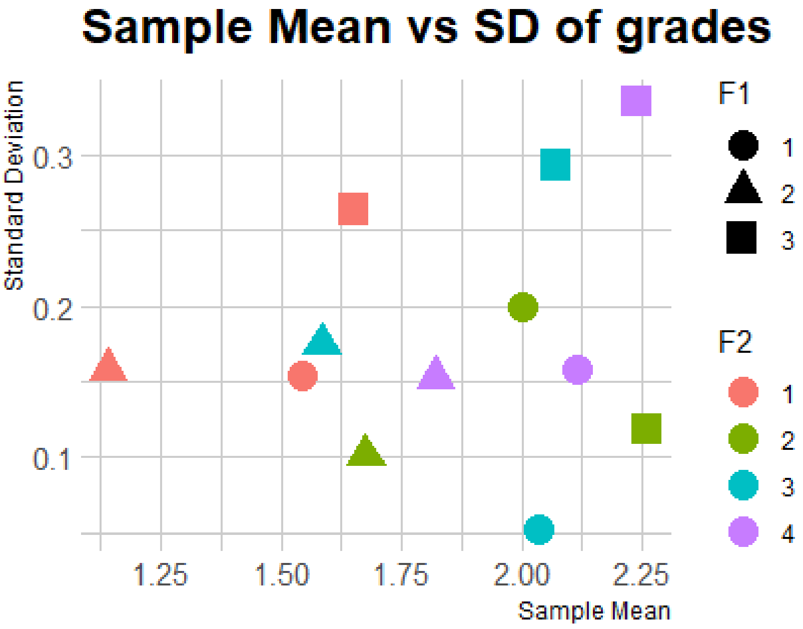
No, we can assume normality of the residuals.

• Do you see any tendency in the residual variance?

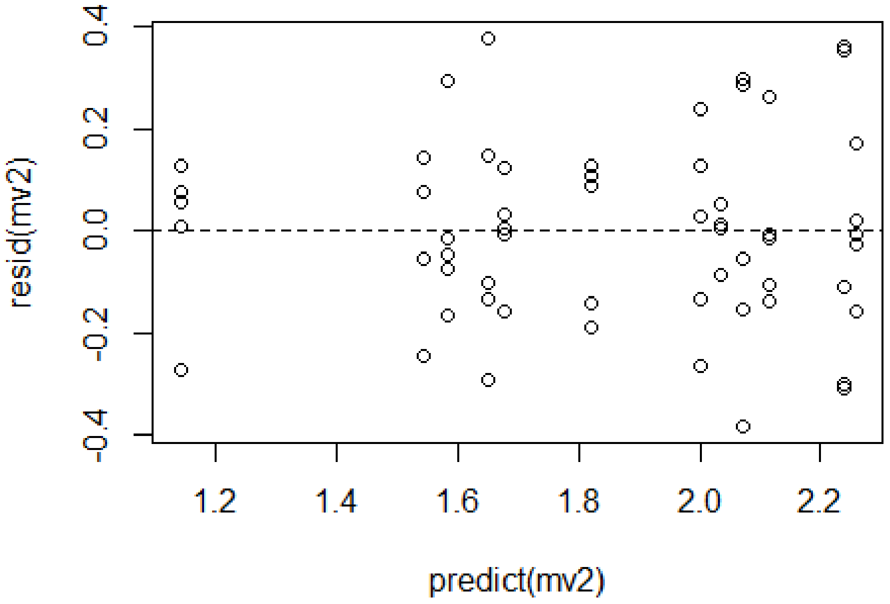
We do not see any abnormal patter in the residuals and can assume homoscedasticity.

1.2. V2

```
mv2<-lm(V2~F1*F2, PRAC2F)
anova(mv2)

## Analysis of Variance Table
##
## Response: V2
##           Df Sum Sq Mean Sq F value    Pr(>F)
## F1         2 2.6846 1.34231 34.8485 4.482e-10 ***
## F2         3 3.3776 1.12588 29.2297 6.728e-11 ***
## F1:F2      6 0.0902 0.01503  0.3902    0.8817
## Residuals 48 1.8489 0.03852
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

a2<-as.vector(with(PRAC2F, tapply(V2, list(F2=F2, F1=F1),mean)))
b2<-as.vector(with(PRAC2F, tapply(V2, list(F2=F2, F1=F1),sd)))

stats<-cbind(a2, b2)
stats<-data.frame(stats)
stats$F1<-as.factor(c(1,1,1,1,2,2,2,2,3,3,3,3))
stats$F2<-as.factor(c(1,2,3,4,1,2,3,4,1,2,3,4))


ggplot(stats, aes(y=b2, x=a2, color=F2, shape=F1) )+ xlab("Standard Deviation
") +ylab("Sample Mean")+ ggtitle("Sample Mean vs SD of grades")+
    geom_point(size=5, ) +
    theme_ipsum()
```



```r
plot(predict(mv2),resid(mv2))
abline(h=0,lty=2)
```



a) Explain a real situation (specifying the name of the variables that appear in the experiment and the procedure to obtain the data) that matches with the given file.

We could think of the response variable to be the crop yield of a farm, one factor the fertilizer we use and the other factor the type of wheat. With this model we would like to pick the best combination of fertilizer and wheat type to maximize our crop yield.

To gather the sufficient data, we should divide our crop in the four types of wheat equally and in each of the wheat type divided it again in three to use each fertilizer in each type of what. At the end of the time selected, e.g. 4 months, we measure the crop yield and analyze it.

• Is there any significant effect? What can be deduced from the means table?

The significant effects are both factors F1 and F2, but have no interaction.

• Which is the error variance estimation?

Mean square error of the residuals: 0.03852.

• Does the first factor have any effect? How is it detected in the means table and in the X – SX diagram?

F1 has effect. We can detect it in the table means by checking how V2 has takes different means when F1 changes from level. If this variability is large enough, F1 has effect on V2.

In the diagram, when checking the effect of F1, we have to observe the different shapes, as each shape represents one level of the first factor. We observe how if we draw a line connecting all the circles, this is not linear, meaning that F1 has an effect on the response variable and the sample means for each level has different value.

• Does the second factor have any effect? How is it detected in the means table and in the X – SX diagram?

F2 has effect. We can detect it in the table means by checking how V2 has takes different means when F2 changes from level. If this variability is large enough, F2 has effect on V2.

In the diagram, when checking the effect of F2, we have to observe the different colors, as each color represents one level of the second factor. We observe how if we draw a line connecting all the blue shapes, this is not linear, meaning that F2 has an effect on the response variable and the sample means for each level has different value.

- **Does it exists interaction? How is it detected in the means table and in the X – SX diagram?**

The interaction is not significant.

In the diagram, to check for interaction, we have to look how the level of each factor gives different means depending on the level of the other factor, this is, how all the circles have a similar value, the same for the squares and triangles. It looks like clusters in the plot.

- **Which is (are) the best treatment?**

The best treatment for V2 is to set the third level of F1 an the second for F2, as it maximizes the expected value of our response variable.

- **Which is (are) the best level of the first factor?**

The third level.

- **Which is (are) the best level of the second factor?**

The fourth level.

- **The best treatment, is it the combination of the best levels of each factor? Why?**

No, in this case, even that we do not detect a significant interaction the combination of the best levels of each factor is not the best treatment, although the expected value is very close, just 0.02 away.

- **If the first level of the first factor needs to be used, which treatment is in your opinion the best?**

If we have to use the first level for F1, then the best treatment would be to use the fourth level for F2, as the expected value is the highest upon the means of the first level of F1, checking the first row of the means table.

- **If the second level of the first factor needs to be used, which treatment is in your opinion the best?**

If we have to use the second level for F1, then the best treatment would be to use the fourth level for F2, as the expected value is the highest upon the means of the second level of F1, checking the second row of the means table.

- **If the third level of the first factor needs to be used, which treatment is in your opinion the best?**

If we have to use the third level for F1, then the best treatment would be to use the second level for F2, as the expected value is the highest upon the means of the third level of F1, checking the third row of the means table.

- **If the first level of the second factor needs to be used, which treatment is in your opinion the best?**

If we have to use the first level for F2, then the best treatment would be to use the third level for F1, as the expected value is the highest upon the means of the first level of F2, checking the first row of the means table.

- **If the second level of the second factor needs to be used, which treatment is in your opinion the best?**

If we have to use the second level for F2, then the best treatment would be to use the third level for F1, as the expected value is the highest upon the means of the second level of F2, checking the second row of the means table.

- **If the third level of the second factor needs to be used, which treatment is in your opinion the best?**

If we have to use the third level for F2, then the best treatment would be to use the third level for F1, as the expected value is the highest upon the means of the third level of F2, checking the third row of the means table.

- **If the fourth level of the second factor needs to be used, which treatment is in your opinion the best?**

If we have to use the fourth level for F2, then the best treatment would be to use the third level for F1, as the expected value is the highest upon the means of the fourth level of F2, checking the fourth row of the means table.

- **Do you see any anomalous tendency in the residuals?**

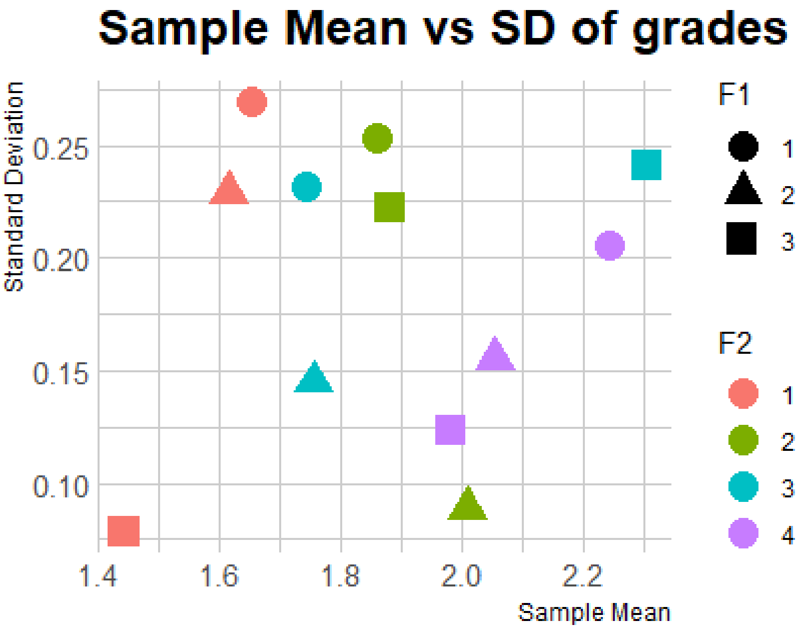
We do not see any abnormal patter in the residuals and can assume homoscedasticity.

1.3. V3

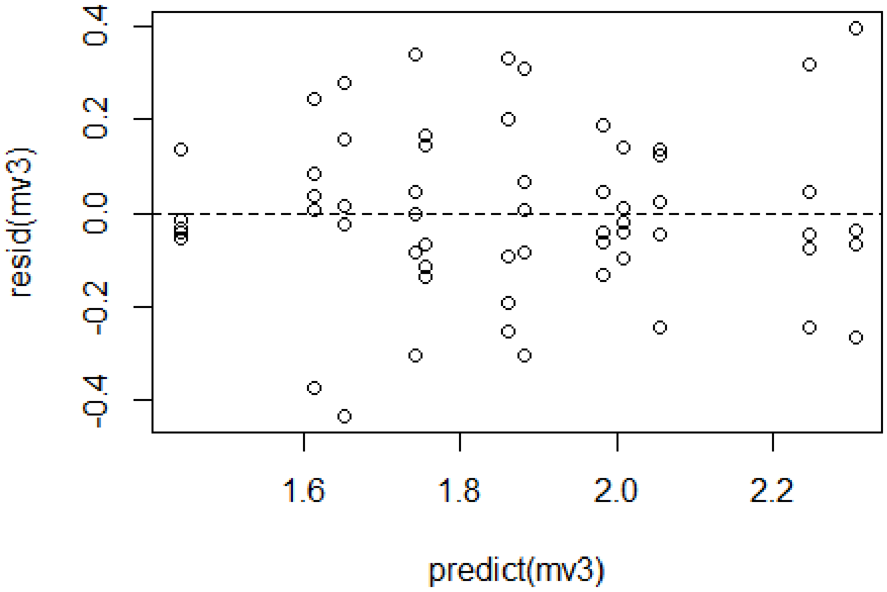
```
mv3<-lm(V3~F1*F2, PRAC2F)
anova(mv3)

## Analysis of Variance Table
##
```

```
## Response: V3
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## F1          2 0.02065 0.01033  0.2657 0.7678019
## F2          3 2.19300 0.73100 18.8071 3.324e-08 ***
## F1:F2       6 1.38236 0.23039  5.9275 0.0001091 ***
## Residuals 48 1.86568 0.03887
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
a3<-as.vector(with(PRAC2F, tapply(V3, list(F2=F2, F1=F1),mean)))
b3<-as.vector(with(PRAC2F, tapply(V3, list(F2=F2, F1=F1),sd)))

stats<-cbind(a3, b3)
stats<-data.frame(stats)
stats$F1<-as.factor(c(1,1,1,1,2,2,2,2,3,3,3,3))
stats$F2<-as.factor(c(1,2,3,4,1,2,3,4,1,2,3,4))


ggplot(stats, aes(y=b3, x=a3, color=F2, shape=F1) )+ xlab("Standard Deviation
") +ylab("Sample Mean")+ ggtitle("Sample Mean vs SD of grades")+
    geom_point(size=5, ) +
    theme_ipsum()
```



```
plot(predict(mv3),resid(mv3))
abline(h=0,lty=2)
```



a) Explain a real situation (specifying the name of the variables that appear in the experiment and the procedure to obtain the data) that matches with the given file.

We can think of the response variable to be the blood pressure of a patient in a hospital and the factors to be the medication type and where from the medication is injected, mouth, arm, toe or butt. This model would help us better understand how to keep blood pressure lower for our patients.

To collect the data we should experiment in 60 patients, with a balanced design we should divide them accordingly and randomly and check their actual blood pressure. After a period of time we should check again the pressure and write down the differences and variation.

• Is there any significant effect? What can be deduced from the means table?

The second factor F2 has a significant effect, as well as the interaction of both factors has significance.

We can see in the mean tables how the overall mean of the levels of F1 does not change much, while the means of F2 do. Similarly we can check that the means of the response variable change by changing the levels of F2, but the values of the levels of F2 depend as well from the levels of F1.

• Which is the error variance estimation?

Mean square error of the residuals: 0.03887.

• Does the first factor have any effect? How is it detected in the means table and in the X – SX diagram?

F1 has no effect on V3.

In the diagram, when checking the effect of F1, we have to observe the different shapes, as each shape represents one level of the first factor. We observe how if we draw a line connecting all the circles, this are linear, meaning that F1 has not an effect on the response variable and the sample means for each level has similar value.

In the diagram, when checking the effect of F2, we have to observe the different colors, as each color represents one level of the second factor. We observe how if we draw a line connecting all the blue shapes, this is not linear, meaning that F2 has an effect on the response variable and the sample means for each level has different value.

• Does the second factor have any effect? How is it detected in the means table and in the X – SX diagram?

F2 has effect. We can detect it in the table means by checking how V3 has takes different means when F2 changes from level. If this variability is large enough, F2 has effect on V3.

In the diagram, to check for interaction, we have to look how the level of each factor gives different means depending on the level of the other factor, this is, how all the circles have a different value, the same for the squares and triangles. We should expect similar values for all the circles if there was no interaction.

• Does it exists interaction? How is it detected in the means table and in the X – SX diagram?

Interaction has a significant effect. In the means table we can detect interaction if we see some significant change if we keep F1 at the same level but change F2. For example, if we look at the means table of V3, we would see how the first level of F1, has different values depending on which level F2 is.

• Which is (are) the best treatment?

The best treatment is level first for F1 and fourth level for F2, as the expected value is the highest among the treatments.

• Which is (are) the best level of the first factor?

The third.

- **Which is (are) the best level of the second factor?**

The fourth.

- **The best treatment, is it the combination of the best levels of each factor? Why?**

No because there is interaction.

- **If the first level of the first factor needs to be used, which treatment is in your opinion the best?**

Second level of F2.

- **If the second level of the first factor needs to be used, which treatment is in your opinion the best?**

Fourth level of F2.

- **If the third level of the first factor needs to be used, which treatment is in your opinion the best?**

Third level of F2.

- **If the first level of the second factor needs to be used, which treatment is in your opinion the best?**

The first level of F1.

- **If the second level of the second factor needs to be used, which treatment is in your opinion the best?**

The second level of F1.

- **If the third level of the second factor needs to be used, which treatment is in your opinion the best?**

The first third of F1.

- **If the fourth level of the second factor needs to be used, which treatment is in your opinion the best?**

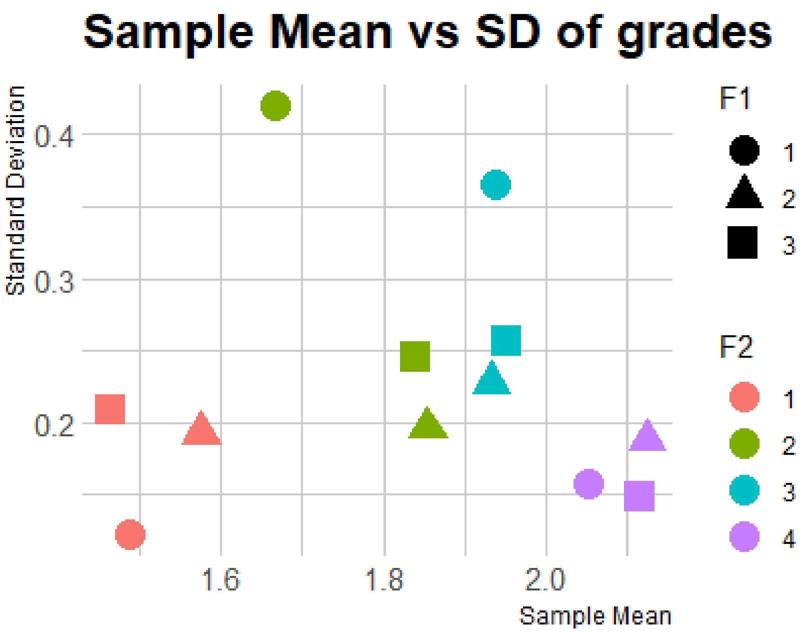
The first level of F1.

- **Do you see any anomalous tendency in the residuals?**

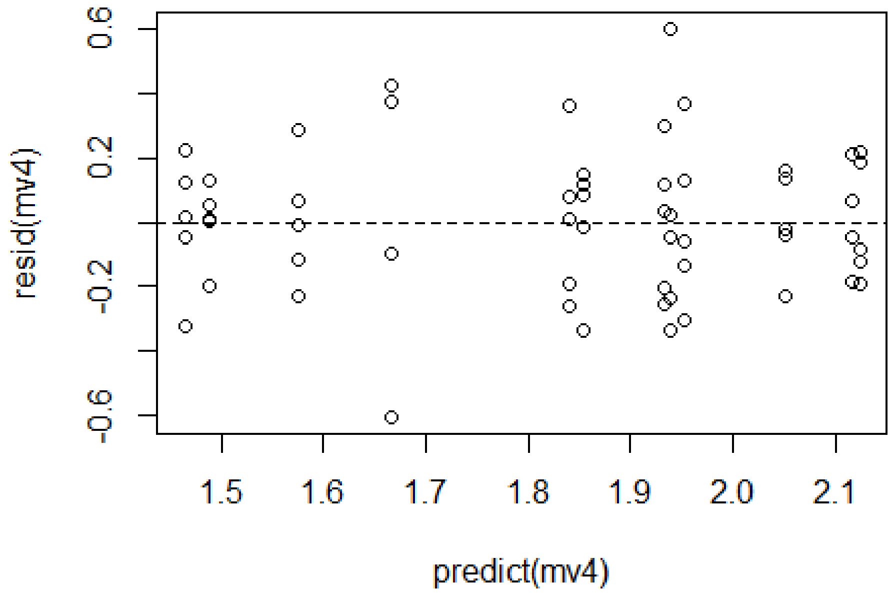
Although there's one treatment that seems to have less variability, we do not see any weird pattern anc can assume homoscedasticity.

1.4. V4

```
mv4<-lm(V4~F1*F2, PRAC2F)
anova(mv4)

## Analysis of Variance Table
##
## Response: V4
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## F1         2 0.07581 0.03790  0.6450    0.5292
## F2         3 2.82555 0.94185 16.0263 2.366e-07 ***
## F1:F2      6 0.08528 0.01421  0.2419    0.9603
## Residuals 48 2.82092 0.05877
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

a4<-as.vector(with(PRAC2F, tapply(V4, list(F2=F2, F1=F1),mean)))
b4<-as.vector(with(PRAC2F, tapply(V4, list(F2=F2, F1=F1),sd)))

stats<-cbind(a4, b4)
stats<-data.frame(stats)
stats$F1<-as.factor(c(1,1,1,1,2,2,2,2,3,3,3,3))
stats$F2<-as.factor(c(1,2,3,4,1,2,3,4,1,2,3,4))


ggplot(stats, aes(y=b4, x=a4, color=F2, shape=F1) )+ xlab("Standard Deviation
") +ylab("Sample Mean")+ ggtitle("Sample Mean vs SD of grades")+
    geom_point(size=5, ) +
    theme_ipsum()
```



```
plot(predict(mv4),resid(mv4))
abline(h=0,lty=2)
```



a) Explain a real situation (specifying the name of the variables that appear in the experiment and the procedure to obtain the data) that matches with the given file.

We can think of the response variable to be the salary of a given person, and the factors to be, his/her race, white, latino, Asian or African and the other to be the football team they support. We want to know how the salary of a person varies depending on his/her race and his favorite football club, Barcelona, Madrid and Athletic Bilbao.

To collect the data we should randomly select 15 people from each race, to get a balanced design we should carefully select them to be even on their football clubs, we have to keep randomly selected. After selecting the observation we measure their yearly salary and complete the table for the analysis.

• Is there any significant effect? What can be deduced from the means table?

Only F2 has a significant effect on V4.

• Which is the error variance estimation?

The mean square error of the residuals: 0.05877.

• Does the first factor have any effect? How is it detected in the means table and in the X – SX diagram?

F1 has no effect, we can check that the overall means of F1 do not vary much.

In the diagram, when checking the effect of F1, we have to observe the different shapes, as each shape represents one level of the first factor. We observe how if we draw a line connecting all the circles, this are linear, meaning that F1 has not an effect on the response variable and the sample means for each level has similar value.

• Does the second factor have any effect? How is it detected in the means table and in the X – SX diagram?

F2 has effect, its overall means from V4 vary a lot depending on the level of F2.

In the diagram, when checking the effect of F2, we have to observe the different colors, as each color represents one level of the second factor. We observe how if we draw a line connecting all the blue shapes, this is not linear, meaning that F2 has an effect on the response variable and the sample means for each level has different value.

• Does it exists interaction? How is it detected in the means table and in the X – SX diagram?

The interaction has no effect, we do not see variation of the factors depending from the other factor.

In the diagram, to check for interaction, we have to look how the level of each factor gives different means depending on the level of the other factor, this is, how all the circles have a similar value, the same for the squares and triangles. It looks like clusters in the plot.

• Which is (are) the best treatment?

The second level for F1 and fourth for F2.

• Which is (are) the best level of the first factor?

The second level.

• Which is (are) the best level of the second factor?

The fourth level.

• The best treatment, is it the combination of the best levels of each factor? Why?

In this case it is, because it does not exist interaction between factors.

• If the first level of the first factor needs to be used, which treatment is in your opinion the best?

The fourth level of F2.

• If the second level of the first factor needs to be used, which treatment is in your opinion the best?

The fourth level of F2.

• If the third level of the first factor needs to be used, which treatment is in your opinion the best?

The fourth level of F2.

• If the first level of the second factor needs to be used, which treatment is in your opinion the best?

The second level of F1.

• If the second level of the second factor needs to be used, which treatment is in your opinion the best?

The second level of F1.

• If the third level of the second factor needs to be used, which treatment is in your opinion the best?

The third level of F1.

• If the fourth level of the second factor needs to be used, which treatment is in your opinion the best?

The second level of F1.

• Do you see any anomalous tendency in the residuals?

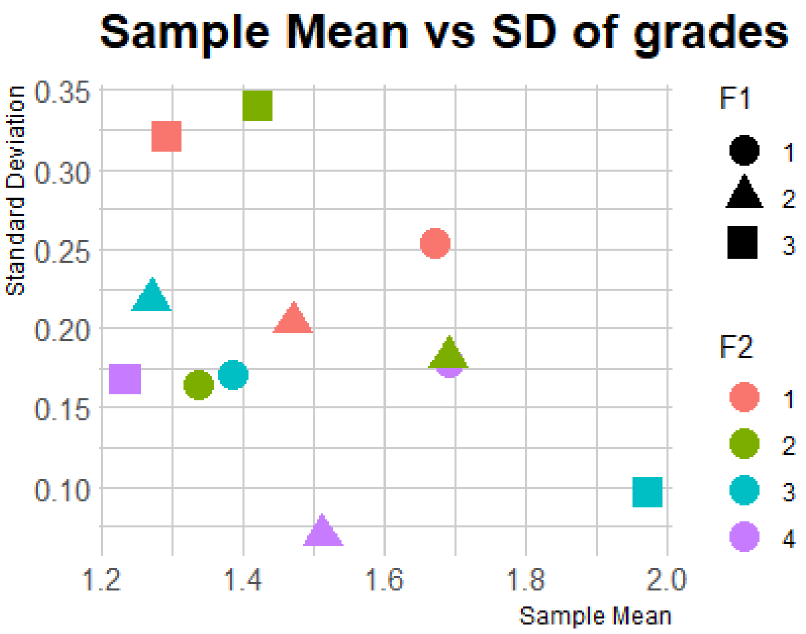
We do not observe any anomaly and can assume homoscedasticity.

1.5. V5

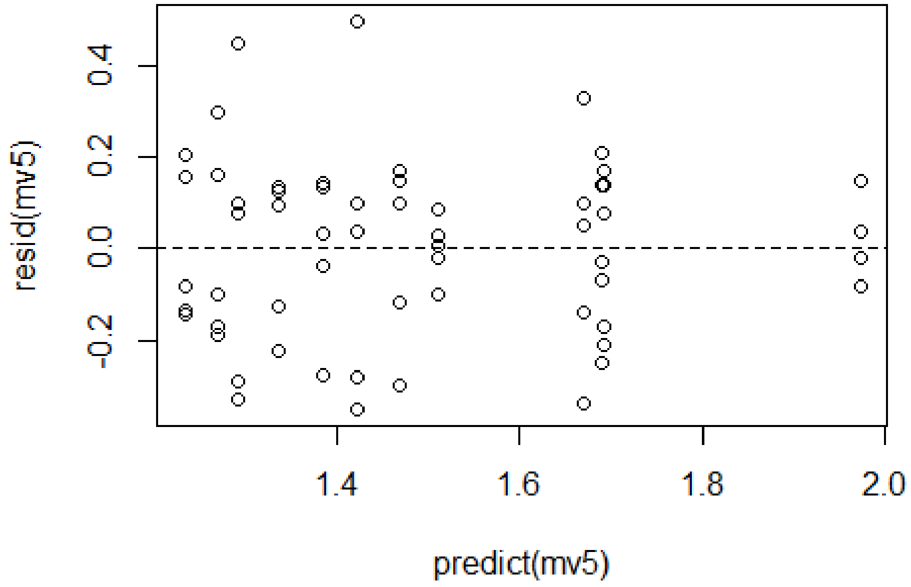
```
mv5<-lm(V5~F1*F2, PRAC2F)
anova(mv5)
```

```
## Analysis of Variance Table
##
## Response: V5
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## F1         2 0.01981 0.00990  0.2219    0.8018
## F2         3 0.04471 0.01490  0.3338    0.8009
## F1:F2      6 2.62720 0.43787  9.8078 4.804e-07 ***
## Residuals 48 2.14296 0.04464
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
a5<-as.vector(with(PRAC2F, tapply(V5, list(F2=F2, F1=F1),mean)))
b5<-as.vector(with(PRAC2F, tapply(V5, list(F2=F2, F1=F1),sd)))

stats<-cbind(a5, b5)
stats<-data.frame(stats)
stats$F1<-as.factor(c(1,1,1,1,2,2,2,2,3,3,3,3))
stats$F2<-as.factor(c(1,2,3,4,1,2,3,4,1,2,3,4))



ggplot(stats, aes(y=b5, x=a5, color=F2, shape=F1) )+ xlab("Standard Deviation
") +ylab("Sample Mean")+ ggtitle("Sample Mean vs SD of grades")+
    geom_point(size=5, ) +
    theme_ipsum()
```



```
plot(predict(mv5),resid(mv5))
abline(h=0,lty=2)
```



a) Explain a real situation (specifying the name of the variables that appear in the experiment and the procedure to obtain the data) that matches with the given file.

We could think of the response variable to be the amount of dirt removed from your laundry and the factors to be the detergent brand used, four different, and the other factor to be the temperature, cold, warm or hot. We ant to know what is the best combination to keep our clothes cleanest.

To collect the data we have to do an experiment, we first write down all the combinations and randomly start the runs for the laundry. Before and after each laundry measure the dirt and compare the dirt removed. Fill the table and we can start to a analyze it.

• Is there any significant effect? What can be deduced from the means table?

Only the interaction has a significant effect on V5.

• Which is the error variance estimation?

Mean square error of the residuals: 0.04464.

• Does the first factor have any effect? How is it detected in the means table and in the X – SX diagram?

F1 has no significant effect, its overall means from V5 do not vary much.

In the diagram, when checking the effect of F1, we have to observe the different shapes, as each shape represents one level of the first factor. We observe how if we draw a line connecting all the circles, this is not linear, meaning that F1 has an effect on the response variable and the sample means for each level has different value.

• Does the second factor have any effect? How is it detected in the means table and in the X – SX diagram?

F2 has no significant effect, its overall means from V5 do not vary much.

In the diagram, when checking the effect of F2, we have to observe the different colors, as each color represents one level of the second factor. We observe how if we draw a line connecting all the blue shapes, this is not linear, meaning that F2 has an effect on the response variable and the sample means for each level has different value.

• Does it exists interaction? How is it detected in the means table and in the X – SX diagram?

Yes the interaction is significant, we can check how the means when taking account each level of each factor vary much.

In the diagram, to check for interaction, we have to look how the level of each factor gives different means depending on the level of the other factor, this is, how all the circles have a similar value, the same for the squares and triangles. It looks like clusters in the plot.

• Which is (are) the best treatment?

The third level of F1 and the third of F2.

• Which is (are) the best level of the first factor?

The first level.

• Which is (are) the best level of the second factor?

The third level.

• The best treatment, is it the combination of the best levels of each factor? Why?

It is not, because it exist interaction and then the values of the levels of the factor depend on the level of the other factor.

• If the first level of the first factor needs to be used, which treatment is in your opinion the best?

The fourth level of F2.

• If the second level of the first factor needs to be used, which treatment is in your opinion the best?

The second level of F2.

• If the third level of the first factor needs to be used, which treatment is in your opinion the best?

The third level of F2.

• If the first level of the second factor needs to be used, which treatment is in your opinion the best?

The first level of F1.

• If the second level of the second factor needs to be used, which treatment is in your opinion the best?

The second level of F1.

- **If the third level of the second factor needs to be used, which treatment is in your opinion the best?**

The third level of F1.

- **If the fourth level of the second factor needs to be used, which treatment is in your opinion the best?**

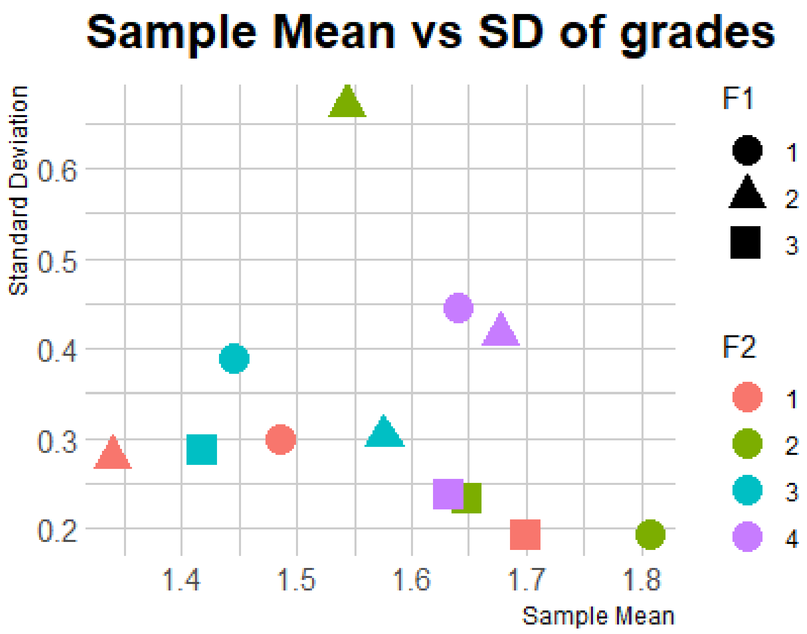
The first level of F1.

- **Do you see any anomalous tendency in the residuals?**

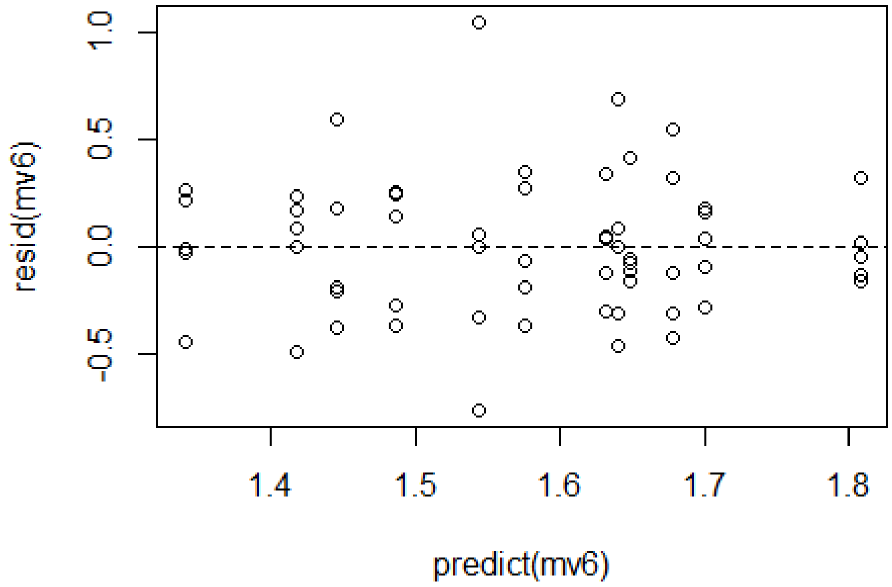
We can presume a megaphone shape and homoscedasticity might not be validated.

1.6. V6

```
mv6<-lm(V6~F1*F2, PRAC2F)
anova(mv6)

## Analysis of Variance Table
##
## Response: V6
##           Df Sum Sq  Mean Sq F value Pr(>F)
## F1         2 0.0527 0.026352  0.2116 0.8100
## F2         3 0.4117 0.137229  1.1021 0.3575
## F1:F2      6 0.5291 0.088187  0.7082 0.6445
## Residuals 48 5.9769 0.124518

a6<-as.vector(with(PRAC2F, tapply(V6, list(F2=F2, F1=F1),mean)))
b6<-as.vector(with(PRAC2F, tapply(V6, list(F2=F2, F1=F1),sd)))

stats<-cbind(a6, b6)
stats<-data.frame(stats)
stats$F1<-as.factor(c(1,1,1,1,2,2,2,2,3,3,3,3))
stats$F2<-as.factor(c(1,2,3,4,1,2,3,4,1,2,3,4))


ggplot(stats, aes(y=b, x=a, color=F2, shape=F1) )+ xlab("Standard Deviation")
+ylab("Sample Mean")+ ggtitle("Sample Mean vs SD of grades")+
    geom_point(size=5, ) +
    theme_ipsum()
```



```r
plot(predict(mv6),resid(mv6))
abline(h=0,lty=2)
```



a) Explain a real situation (specifying the name of the variables that appear in the experiment and the procedure to obtain the data) that matches with the given file.

We can think of the response variable to be the time to detect a tumor, and the factors are four different doctors, James, Paul, Susan and Sophie and the other one to be the methodology, three different methods to check for tumors. With this model we want to know if there are differences between doctors and if the methodology used has any effect on detecting faster the tumor.

To collect the data we need the medical records of the four doctors and select a balanced design with the different kind of methodologies. After fitting the table we can start our analysis.

• Is there any significant effect? What can be deduced from the means table?

There are no significant effects. The overall means of the factors do not vary much from level to level.

• Which is the error variance estimation?

The mean square error of the residuals: 0.124518.

• Does the first factor have any effect? How is it detected in the means table and in the X – SX diagram?

F1 has no significant effect, F1 overall means do not vary much.

In the diagram, when checking the effect of F1, we have to observe the different shapes, as each shape represents one level of the first factor. We observe how if we draw a line connecting all the circles, this are linear, meaning that F1 has not an effect on the response variable and the sample means for each level has similar value.

• Does the second factor have any effect? How is it detected in the means table and in the X – SX diagram?

F2 has no significant effect, F2 overall means do not vary much.

In the diagram, when checking the effect of F2, we have to observe the different colors, as each color represents one level of the second factor. We observe how if we draw a line connecting all the blue shapes, this is linear, meaning that F2 has not an effect on the response variable and the sample means for each level has similar value.

• Does it exists interaction? How is it detected in the means table and in the X – SX diagram?

There is no interaction, we do not see variation in the means by switching levels.

In the diagram, to check for interaction, we have to look how the level of each factor gives different means depending on the level of the other factor, this is, how all the circles have a similar value, the same for the squares and triangles. It looks like clusters in the plot.

- **Which is (are) the best treatment?**

First level of F1 and second level of F2, as the expected value is the highest.

- **Which is (are) the best level of the first factor?**

The third level.

- **Which is (are) the best level of the second factor?**

The second level.

- **The best treatment, is it the combination of the best levels of each factor? Why?**

No, in this case there is no interaction, but there is no significant relation between the factors and V5, so it does not matter the best treatment because is just random.

- **If the first level of the first factor needs to be used, which treatment is in your opinion the best?**

It does not matter because the relation is not significant.

- **Do you see any anomalous tendency in the residuals?**

There is some anomaly, and we cannot assume homoscedasticity.

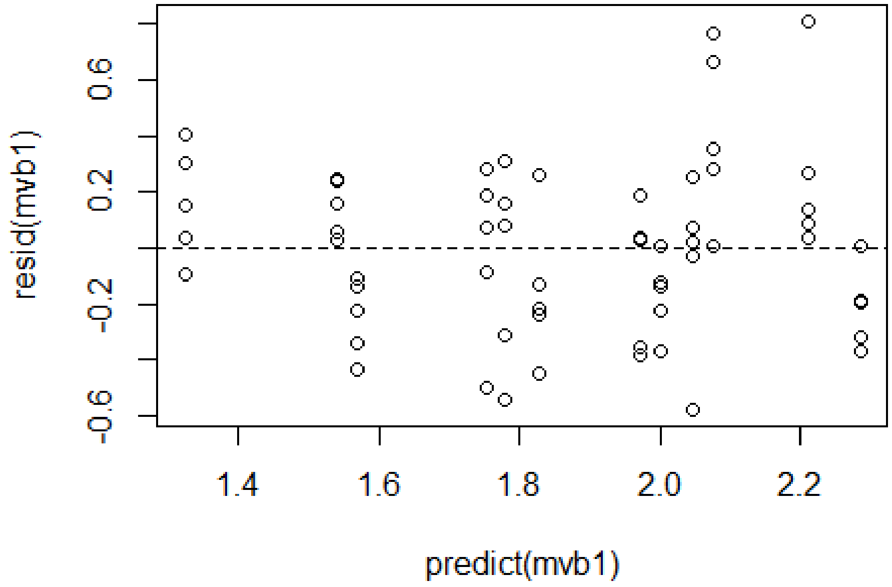
Conclusion: This exercise helps us to understand what the interaction effect means and how to check it by means of the means table and means diagram. We have checked how interaction means that the means of a response variable depends on both the levels of the factors, even if the factors alone are not significantly related to the response variable

2. Additive models:

We assume this model: $Y_{ij} = \beta_0 + \beta_1 * F1_i + \beta_2 * F2_j$. We will fit the data for each response variable.

2.2. V1

```
mvb1<-lm(V1~F1+F2, PRAC2F)
anova(mvb1)

## Analysis of Variance Table
##
## Response: V1
##           Df Sum Sq Mean Sq F value    Pr(>F)
## F1         2 2.9665 1.48323 15.0289 6.472e-06 ***
## F2         3 1.5610 0.52035  5.2724  0.002909 **
## Residuals 54 5.3294 0.09869
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

plot(predict(mvb1),resid(mvb1))
abline(h=0,lty=2)
```



- **Is there any significant effect?**

Yes both factors have a significant effect on the response variable.

- **Do you see any anomalous tendency in the residuals?**

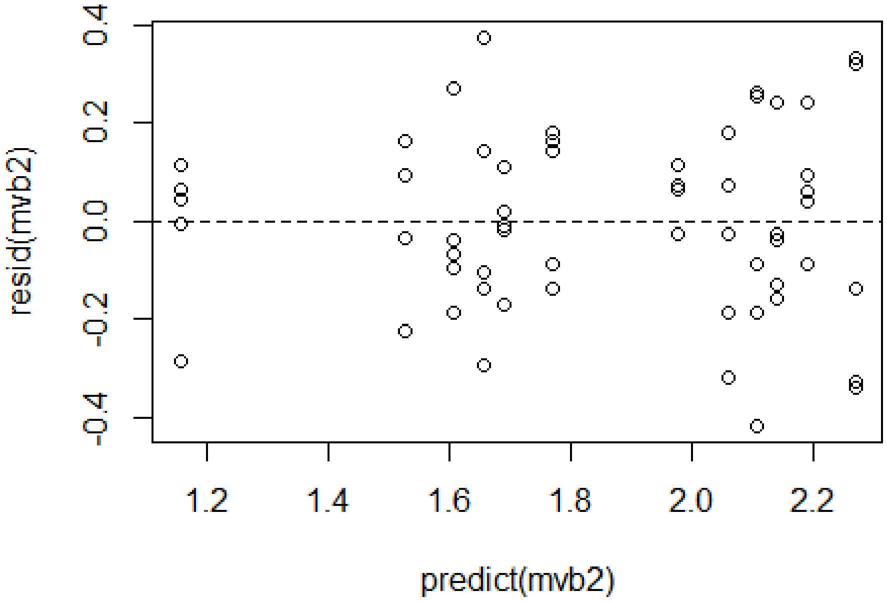
Although we do not see too much linearity in the residuals we do not see a weird patter, so we assume homoscedasticity.

- **Do you think that the two-way ANOVA is an appropriate model for these data?**

Yes, in this case both factors are significant, so we have to include them both in our model.

2.2. V2

```
mvb2<-lm(V2~F1+F2, PRAC2F)
anova(mvb2)

## Analysis of Variance Table
##
## Response: V2
##           Df Sum Sq Mean Sq F value    Pr(>F)
## F1         2 2.6846 1.34231  37.381 6.461e-11 ***
## F2         3 3.3776 1.12588  31.354 7.122e-12 ***
## Residuals 54 1.9391 0.03591
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

plot(predict(mvb2),resid(mvb2))
abline(h=0,lty=2)
```



- **Is there any significant effect?**

Both factors are significant to the response variable.

- **Do you see any anomalous tendency in the residuals?**

No, we can assume homocedasticity.

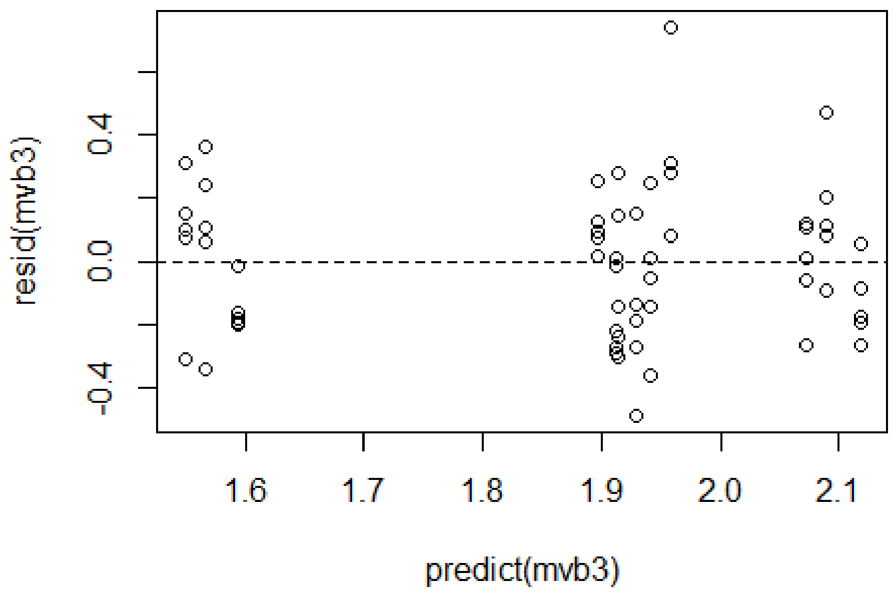
- **Do you think that the two-way ANOVA is an appropriate model for these data?**

Yes, in this case both factors are significant, so we have to include them both in our model.

2.3. V3

```
mvb3<-lm(V3~F1+F2, PRAC2F)
anova(mvb3)

## Analysis of Variance Table
##
## Response: V3
##           Df Sum Sq Mean Sq F value    Pr(>F)
## F1         2 0.0207 0.01033  0.1717    0.8427
## F2         3 2.1930 0.73100 12.1532 3.453e-06 ***
## Residuals 54 3.2480 0.06015
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(predict(mvb3),resid(mvb3))
abline(h=0,lty=2)
```



- **Is there any significant effect?**

Only F2 is significant to the response variable.

- **Do you see any anomalous tendency in the residuals?**

We see some potential outlier, but overall we can assume homoscedasticity.

- **Do you think that the two-way ANOVA is an appropriate model for these data?**

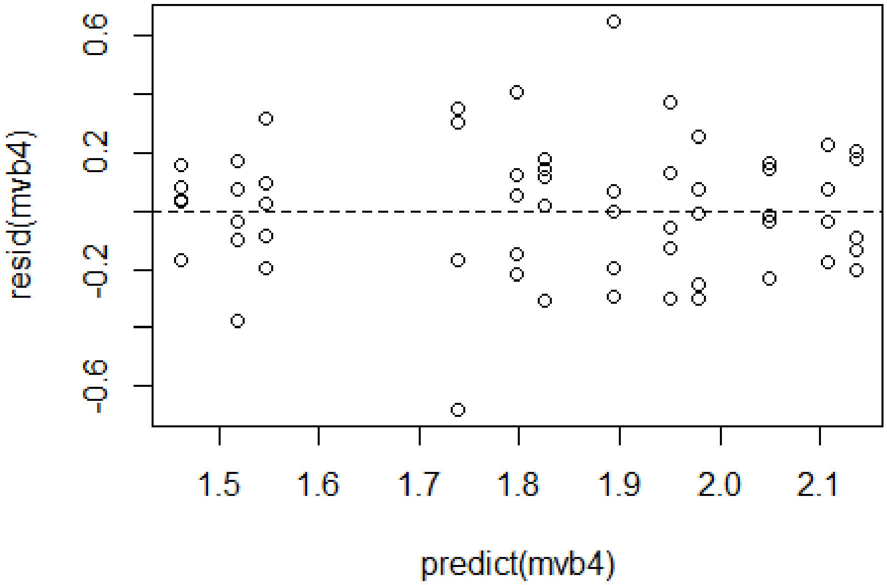
In this case, as only one factor is significant, we could get rid of the other factor, and with less parameters our model would be better, a one-way anova.

2.4. V4

```
mvb4<-lm(V4~F1+F2, PRAC2F)
anova(mvb4)
```

```
## Analysis of Variance Table
##
## Response: V4
##           Df  Sum Sq Mean Sq F value    Pr(>F)    
## F1         2 0.07581 0.03790  0.7043    0.4989    
## F2         3 2.82555 0.94185 17.5005 4.616e-08 ***
## Residuals 54 2.90620 0.05382                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(predict(mvb4),resid(mvb4))
abline(h=0,lty=2)
```



• Is there any significant effect?

Only F2 has a significant effect on the response variable.

• Do you see any anomalous tendency in the residuals?

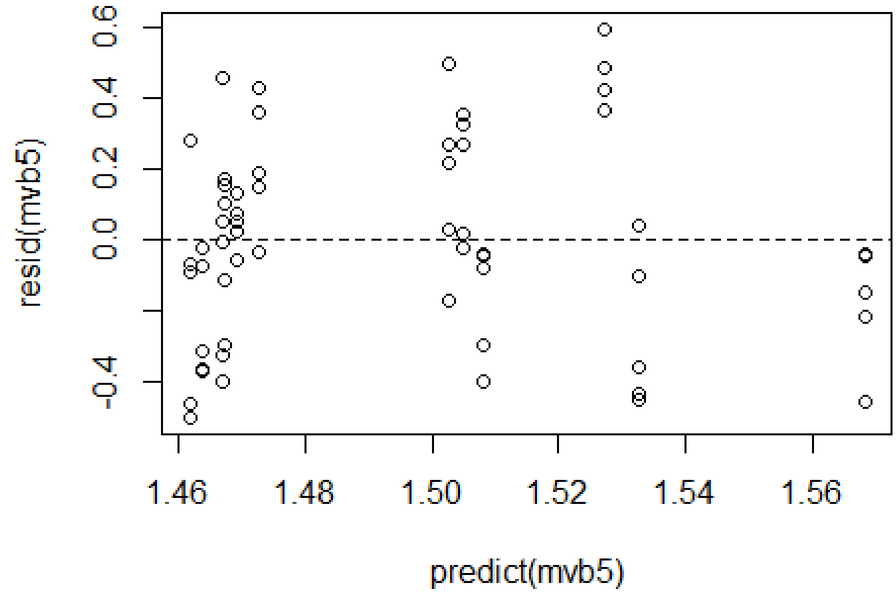
We do not see any anomaly in the residuals, we can assume homoscedasticity.

• Do you think that the two-way ANOVA is an appropriate model for these data?

In this case, as only one factor is significant, we could get rid of the other factor, and with less parameters our model would be better, a one-way anova.

2.5. V5

```
mvb5<-lm(V5~F1+F2, PRAC2F)
anova(mvb5)

## Analysis of Variance Table
##
## Response: V5
##           Df Sum Sq  Mean Sq F value Pr(>F)
## F1         2 0.0198 0.009905  0.1121 0.8941
## F2         3 0.0447 0.014904  0.1687 0.9170
## Residuals 54 4.7702 0.088336

plot(predict(mvb5),resid(mvb5))
abline(h=0,lty=2)
```



• Is there any significant effect?

None of the factors is significant.

• Do you see any anomalous tendency in the residuals?

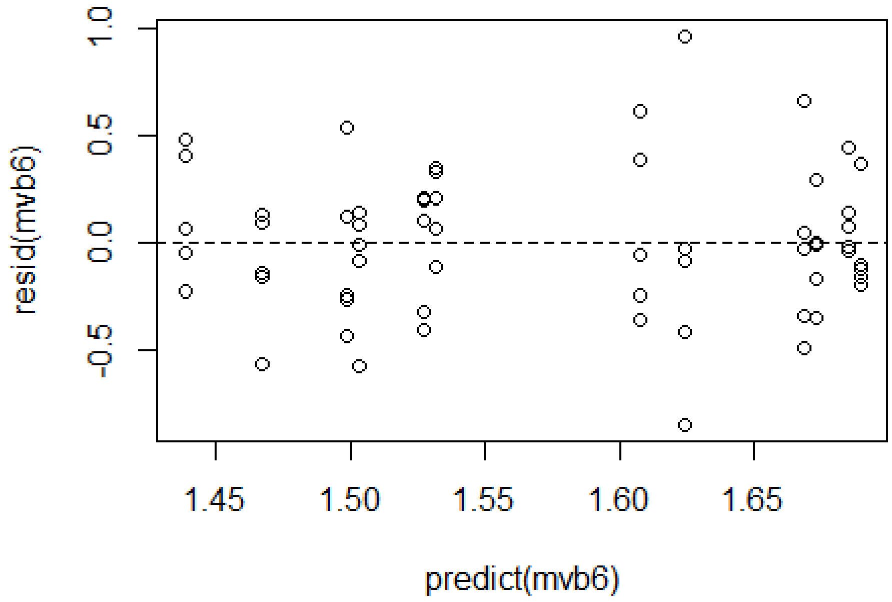
There is some anomaly in the residuals, but as the model does not explain anything, it does not matter.

• Do you think that the two-way ANOVA is an appropriate model for these data?

In this case, our factors are not significant, so we cannot predict or explain the response variable variability.

2.6. V6

```
mvb6<-lm(V6~F1+F2, PRAC2F)
anova(mvb6)

## Analysis of Variance Table
##
## Response: V6
##           Df Sum Sq  Mean Sq F value Pr(>F)
## F1         2 0.0527 0.026352  0.2187 0.8043
## F2         3 0.4117 0.137229  1.1390 0.3416
## Residuals 54 6.5060 0.120482

plot(predict(mvb6),resid(mvb6))
abline(h=0,lty=2)
```



- **Is there any significant effect?**

None of the factors is significant.

- **Do you see any anomalous tendency in the residuals?**

There is some anomaly in the residuals, but as the model does not explain anything, it does not matter.

- **Do you think that the two-way ANOVA is an appropriate model for these data?**

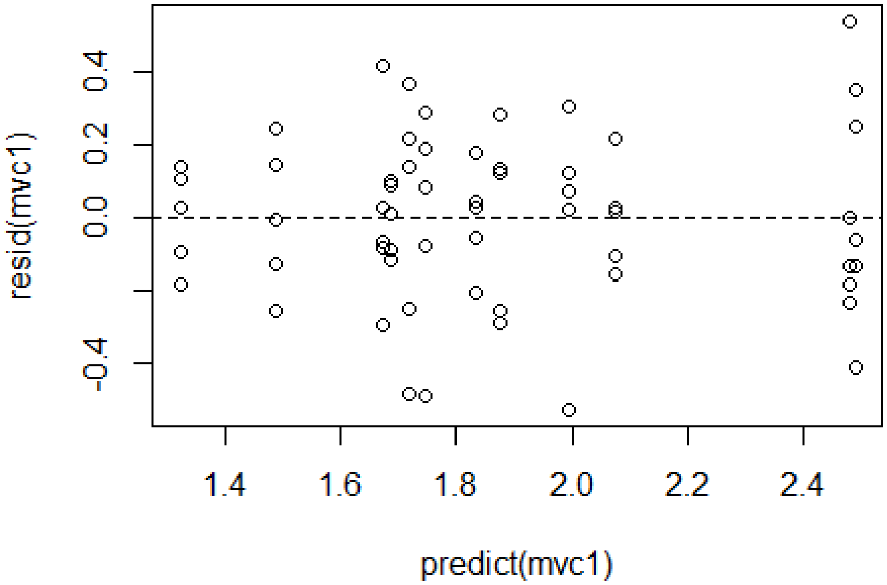
In this case, our factors are not significant, so we cannot predict or explain the response variable variability.

3. Nested models:

We assume this model: $Y_{ij} = \beta_0 + \beta_1 * F1_i + (\beta_3 * F1_i * F2_j)$. Only the first factor and the interaction term.

3.1. V1

```
mvc1<-lm(V1~F1+F1:F2, PRAC2F)
anova(mvc1)

## Analysis of Variance Table
##
## Response: V1
##           Df Sum Sq Mean Sq F value    Pr(>F)
## F1         2 2.9665 1.48323 23.7530 6.746e-08 ***
## F1:F2      9 3.8931 0.43257  6.9272 2.534e-06 ***
## Residuals 48 2.9973 0.06244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

plot(predict(mvc1),resid(mvc1))
abline(h=0,lty=2)
```



- **Is there any significant effect?**

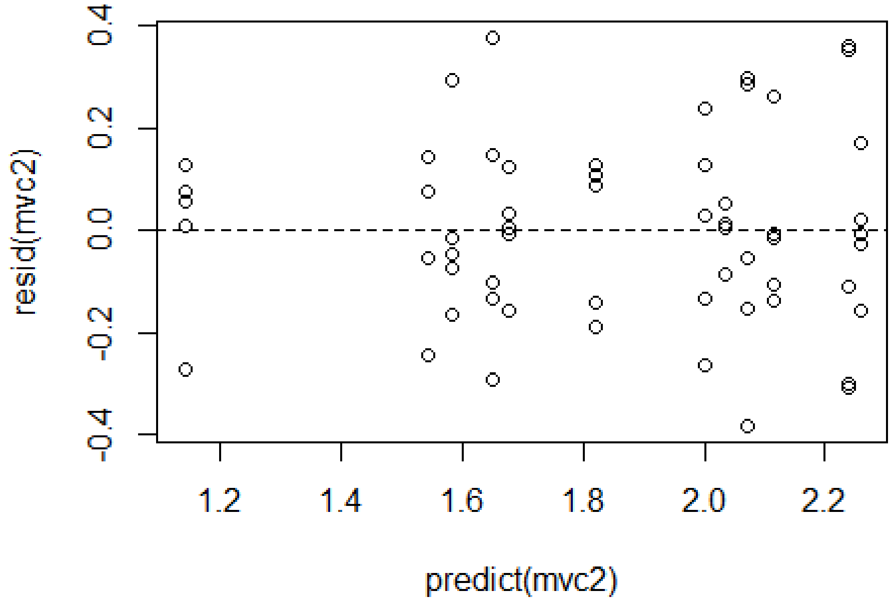
Both factors are significant to the response variable.

- **Do you see any anomalous tendency in the residuals?**

We see a kind of a megaphone shape, so we do not assume homoscedasticity.

3.2. V2

```
mvc2<-lm(V2~F1+F1:F2, PRAC2F)
anova(mvc2)

## Analysis of Variance Table
##
## Response: V2
##           Df Sum Sq Mean Sq F value    Pr(>F)
## F1         2 2.6846 1.34231  34.849 4.482e-10 ***
## F1:F2      9 3.4678 0.38531  10.003 1.879e-08 ***
## Residuals 48 1.8489 0.03852
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

plot(predict(mvc2),resid(mvc2))
abline(h=0,lty=2)
```



- **Is there any significant effect?**

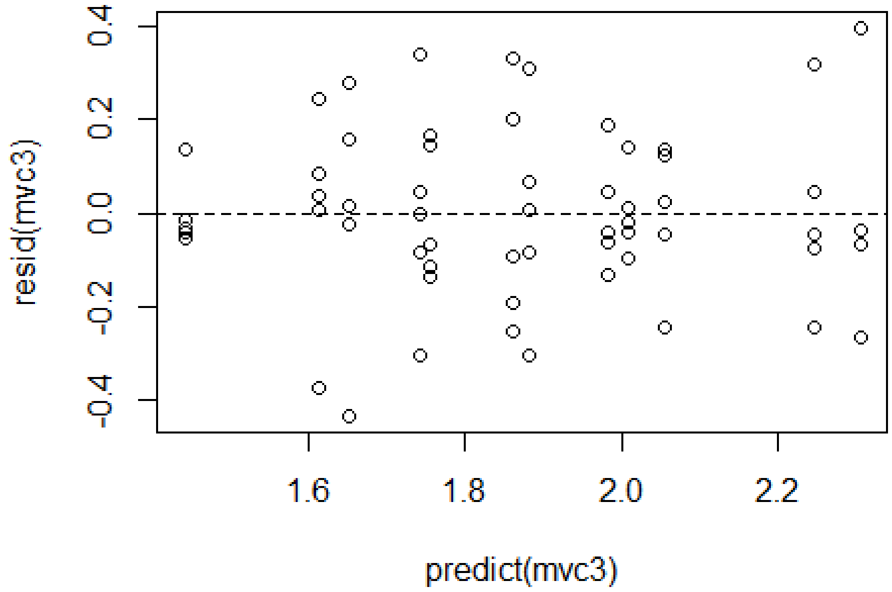
Both factors are significant to the response variable.

- **Do you see any anomalous tendency in the residuals?**

We see a kind of a megaphone shape, so we do not assume homoscedasticity.

3.3. V3

```
mvc3<-lm(V3~F1+F1:F2, PRAC2F)
anova(mvc3)

## Analysis of Variance Table
##
## Response: V3
##           Df Sum Sq Mean Sq F value    Pr(>F)
## F1         2 0.0207 0.01033  0.2657    0.7678
## F1:F2      9 3.5754 0.39726 10.2207 1.374e-08 ***
## Residuals 48 1.8657 0.03887
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

plot(predict(mvc3),resid(mvc3))
abline(h=0,lty=2)
```



- **Is there any significant effect?**

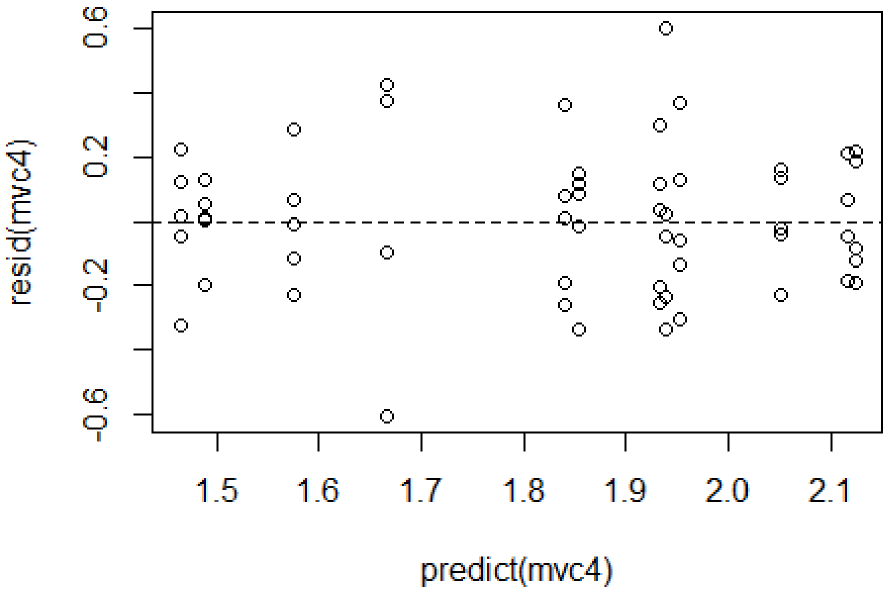
Only the interaction has a significant effect on the response variable

- **Do you see any anomalous tendency in the residuals?**

Although in the beginning the variance seem lower than the rest, overall we can assume homoscedasticity.

3.4. V4

```
mvc4<-lm(V4~F1+F1:F2, PRAC2F)
anova(mvc4)

## Analysis of Variance Table
##
## Response: V4
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## F1         2 0.07581 0.03790  0.6450    0.5292
## F1:F2      9 2.91083 0.32343  5.5033 3.476e-05 ***
## Residuals 48 2.82092 0.05877
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

plot(predict(mvc4),resid(mvc4))
abline(h=0,lty=2)
```



- **Is there any significant effect?**

Only the interaction has a significant effect on the response variable

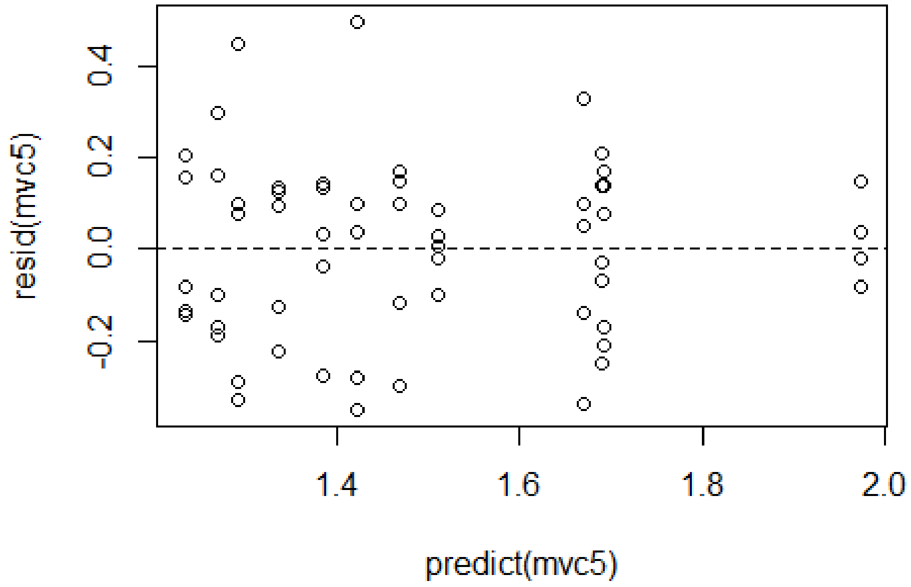
- **Do you see any anomalous tendency in the residuals?**

Overall we can assume homoscedasticity.

3.5. V5

```
mvc5<-lm(V5~F1+F1:F2, PRAC2F)
anova(mvc5)

## Analysis of Variance Table
##
## Response: V5
##           Df  Sum Sq  Mean Sq F value    Pr(>F)
## F1         2 0.01981 0.009905  0.2219    0.8018
## F1:F2      9 2.67191 0.296879  6.6498 4.141e-06 ***
## Residuals 48 2.14296 0.044645
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(predict(mvc5),resid(mvc5))
abline(h=0,lty=2)
```



- **Is there any significant effect?**

Only the interaction has a significant effect on the response variable

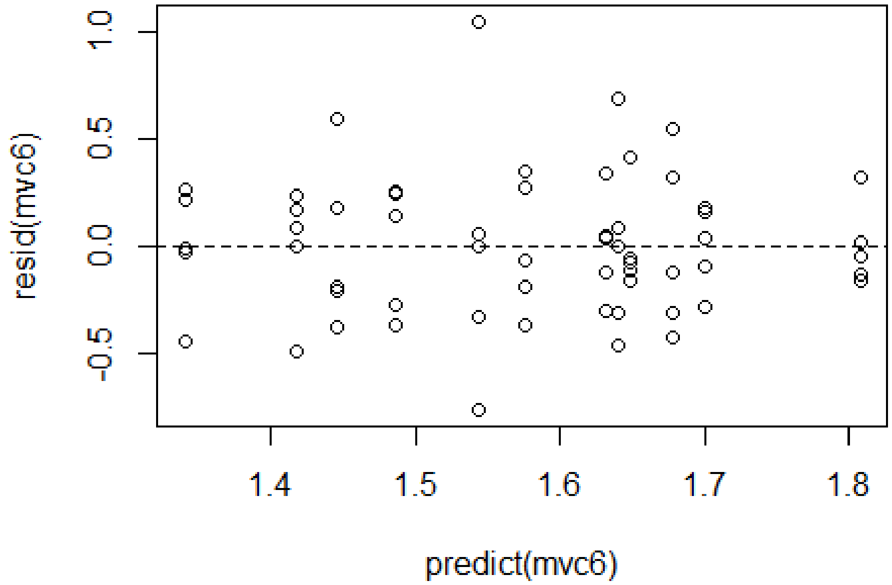
- **Do you see any anomalous tendency in the residuals?**

Although in the end the variance seem lower than the rest, overall we can assume homoscedasticity.

3.6. V6

```
mvc6<-lm(V6~F1+F1:F2, PRAC2F)
anova(mvc6)

## Analysis of Variance Table
##
## Response: V6
##           Df Sum Sq  Mean Sq F value Pr(>F)
## F1         2 0.0527 0.026352  0.2116 0.8100
## F1:F2      9 0.9408 0.104534  0.8395 0.5839
## Residuals 48 5.9769 0.124518
```

```
plot(predict(mvc6),resid(mvc6))
abline(h=0,lty=2)
```



- **Is there any significant effect?**

F1 and the interaction are not significant.

- **Do you see any anomalous tendency in the residuals?**

Homoscedasticity seems to be not verified.

Conclusions: We have seen three different designs for the same variables, the factorial, the additive and the nested design. It is important to understand all of them in order to best choose the model for our data. Depending on what parameters are significant, each model adjusts better to the data.