Assignment 3

Code

Magnus S Grytten

Hide

```
library(FrF2)
```

Issues to be adressed

In this project I have studied how different factors; age, gender and high school grades, effects norwegian spelling. This is not something I a lot of knowledge about, but my suspicion is that spelling will improve with age and grades. Therefore, this is an intersting project to look into and it lets me look futher into Two-level factorial designs. I want to acheve a better understanding of how different factors influence how well people spell, and learn about Design of experiments.

Selection of factors and levels

My hypothesis is that the person' age and what grade the person got norwegian final year of highshool, is going to affect how good the person is at spelling. I am also curious to how gender is going to affect it, even though i imagine it wont have a large impact. How much a person reads and writes in their work and daily life probably also affects a persons spelling ability. However those factors would be harder to precisely quantify and would have to be self reported by the people participating in the test. It would be hard to control that those facors were at the desired level. Therefore the factors im going to look at are gender, age, and norwegian grade.

I dont expect to see interaction between the factors,as i would think they independently affect the response. There isnt much reasoning behind this exept gut feeling.

For gender the levels are obviusly man and woman. For grades i chose the grade 5 for the high level and 4 for the low level. I have to consider who I can get to take part in my experiment. I therefore I chose theese grades as, they are the most common grades for people I know. When it comes to age i need both levels to be above at least 19 as the test subjects need to have a final grade from highshool. For the low level i chose 21-22 years as thats the age most of my friends are. For the high level i choose the ages of 45-50 years. Theese age groups are far enough apart that I think we are going to find a effect if its there.

Controlling that levels are at the desired level is easy, as age, gender and grades aren't going to change for the duration of the experiment, and are either obvius or easely checked. I will however have to trust that the people taking the test are truthfull about, particularly their grades.

Selection of response variable

The response variable that provides information about the problem is the probability for a person to spell a random word incorrectly. The responce is going to be mesured by how many words the test person spells incorrectly in a spellingtest, devided by the total ammount of words in the test. The spelling test is going to be comprised of more difficult words to spell, since easy words that almost any norwegian adult is going to be able to spell probobly wont add any information. As most likly all the test subjects would write them corretly. Thus if the test was made up by completly random words it would have to be very long to be able to find any statistical significance. This does however mean that the model wont do a good job indicating the probability for a person to spell any word correctly. Rather it gives more of a "qualitative" prediction for a persons spelling ability. The accuracy of the mesurement is going to be exact, as its just the amount of incorrectly spelled words. But even

though the mesurment is exact this doesnt mean that the test gives the exact probability for the person to spell any word, or even the words in the test incorrectly. As people spell the same words both correct and incorrect all the time.

Choice of design

My experiment is going to be a 2^3 experiment, looking at three facors at two levels. From earlier i found three factors wich were relevant and easy to controll for. I am also going to want do do the experiment in replicates. I suspect there is going to be a high variance in the response. By replicating the experiment the aproximator for the effects is going to be more accurate. Since i am going to do replicates, doing a using a fewer factors or using a fractional factorial design is going to be usefull to keep the amount of observations lower. A blocked design isnt necessary as all of the experiments are in efect beeing done at the same time (I send out the test everyone at the same time) so the experimental contitions wont realy change for the duration of the tests, in adition its unliky that any changes in conditions would infuence the responce values.

Implementation of the experiment

Randomization is naturaly followed, since the test is being done in the random order of people opening their email, by people in completly different locations. The chances of external factors influencing the responce in a way that skewes the results is low. Each experiment is a genuine run replicate. Each trial is preformed independently and is a full trial.

Analysis of data

Theese are the results, were we have:

$$(+, -)$$

$$A = Gender(Male, Female)$$

$$B = Age(21 - 22, 45 - 50)$$

$$C = Grade(4, 5)$$

Hide

```
print(Resultater)
```

A <dbl>	B <dbl>	C <dbl>	y1 <dbl>	y2 <dbl>
-1	-1	-1	0.18750	0.37500
1	-1	-1	0.40625	0.50000
-1	1	-1	0.46875	0.34375
1	1	-1	0.37500	0.28125
-1	-1	1	0.18750	0.34375
1	-1	1	0.31250	0.40625
-1	1	1	0.09375	0.31250
1	1	1	0.03125	0.18750

8 rows

Calculating the effects:

I use $\frac{y_1+y_2}{2}$ as the y value since i have replicates. This estimates the same effects as $\frac{\sum_i \partial_i Y_i}{n/2}$ would do.

Hide

```
results <- Resultater[,1:3]
results$y <- (Resultater$y1+Resultater$y2)/2
lm3 <- lm.default(formula = y ~ (.)^3, data = results)
summary(lm3)
```

```
Call:
lm.default(formula = y ~ (.)^3, data = results)

Residuals:
ALL 8 residuals are 0: no residual degrees of freedom!

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.300781         NA      NA       NA
A            0.011719         NA      NA       NA
B           -0.039062         NA      NA       NA
C           -0.066406         NA      NA       NA
A:B         -0.054688         NA      NA       NA
A:C         -0.011719         NA      NA       NA
B:C         -0.039062         NA      NA       NA
A:B:C        0.007813         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:    NaN
F-statistic:   NaN on 7 and 0 DF,  p-value: NA
```

Hide

```
effects <- 2*lm3$coeff
print(effects)
```

```
(Intercept)           A           B           C         A:B         A:C
  0.6015625   0.0234375  -0.0781250  -0.1328125  -0.1093750  -0.0234375
        B:C       A:B:C
 -0.0781250   0.0156250
```

Hide

```
#anova(lm3)
```

Making sure the estemated effects are corect by checking the effect of A

Hide

```
M <- data.matrix(Resultater)
i <- 1
print((M[,i]%*%M[,4]+M[,i]%*%M[,5])/8)
```

```
         [,1]
[1,] 0.0234375
```

Checking for statistical significance: Using a significance level of $a = 0.05$ we get,

$$t_{\frac{a}{2},8} = 2.306$$

since we have 8 degrees of freedom.

Using $s^2 = \frac{\sum_i \frac{(y_{1i}-y_{2i})^2}{2}}{n/2}$, and $s^2_{effect} = \frac{4\sigma}{n}$ we get,

Hide

```
s2 <- mean(((M[,4]-M[,5])^2)/2)
sig2 <- 4*s2/16
print(sig2)
```

```
[1] 0.002716064
```

Hide

```
t_025_8 <- 2.306
```

Checking the significance of the effects by checking if $|effect| \geq t_{\frac{a}{2},8} * s^2_{effect}$ for each effect.

Hide

```
print(sig2*t_025_8)
```

```
[1] 0.006263245
```

Hide

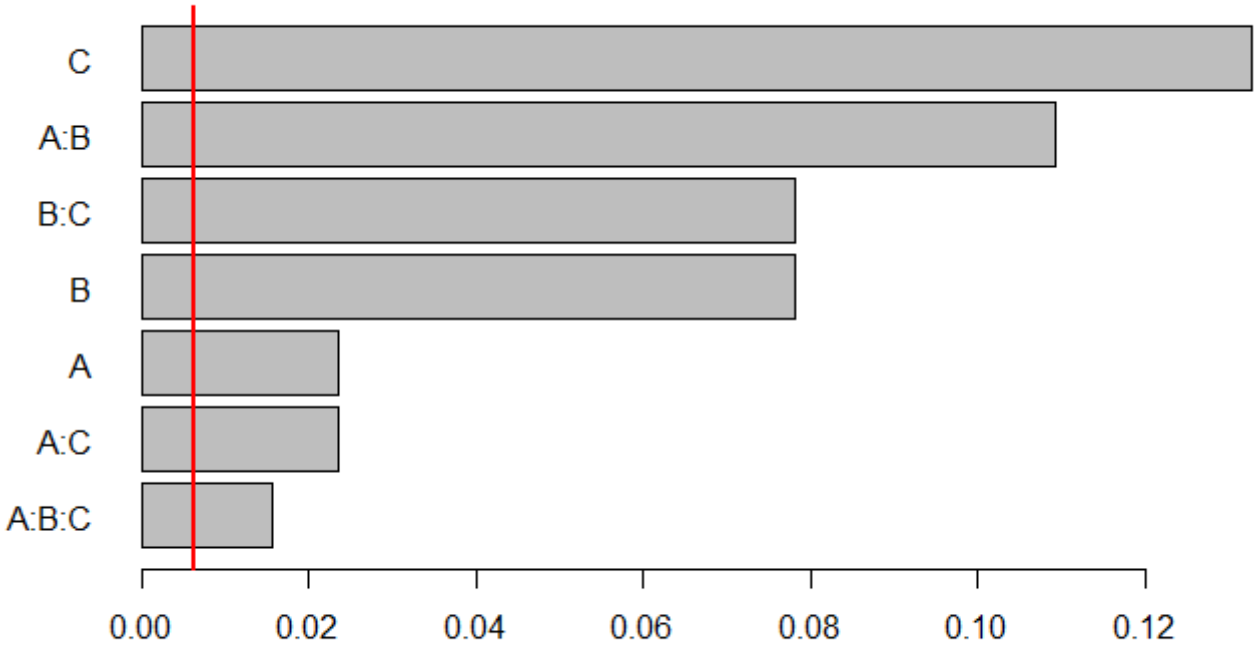
```
print(abs(effects)>sig2*t_025_8)
```

```
(Intercept)           A           B           C         A:B         A:C
       TRUE        TRUE        TRUE        TRUE        TRUE        TRUE
        B:C       A:B:C
       TRUE        TRUE
```

Ploting the test:

Hide

```
barplot(sort((abs(effects[-1])),decreasing=FALSE),las=1,horiz=TRUE,cex.names=1.0)
abline(v=sig2*t_025_8,col=2,lwd=2)
```

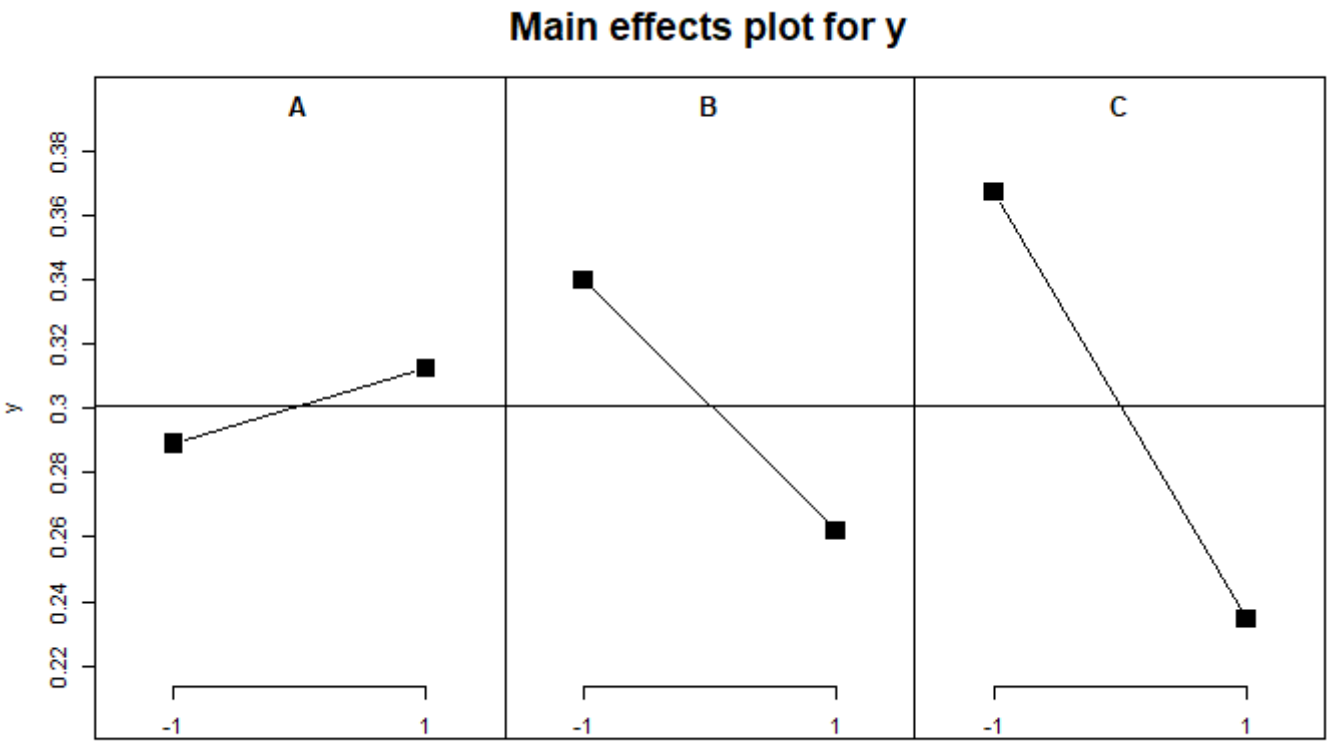


We find that every effect and interaction is statisticaly significant

Ploting the main effects and the interactions:

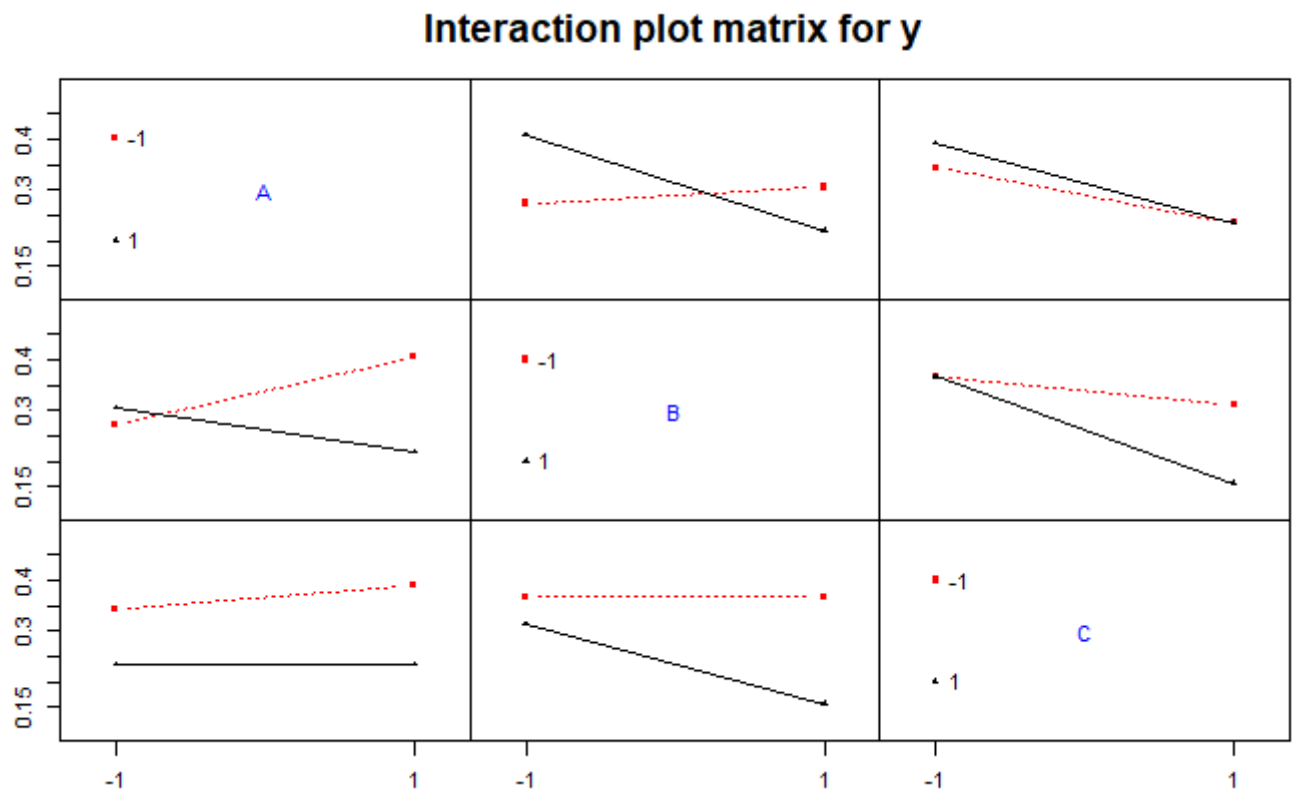
Hide

```
# main and interaction effects plots
MEPlot(lm3)
```



Hide

```
IAPlot(lm3)
```



Checking residuals for a reduced model with main effects and first order interactions

Hide

```
lm2 <- lm.default(formula = y ~  (.)^2, data = results)
summary(lm2)
```

```
Call:
lm.default(formula = y ~ (.)^2, data = results)

Residuals:
        1         2         3         4         5         6         7         8
-0.007813  0.007813  0.007813 -0.007813  0.007813 -0.007813 -0.007813  0.007813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.300781   0.007813    38.5   0.0165 *
A            0.011719   0.007813     1.5   0.3743
B           -0.039062   0.007813    -5.0   0.1257
C           -0.066406   0.007813    -8.5   0.0746 .
A:B         -0.054688   0.007813    -7.0   0.0903 .
A:C         -0.011719   0.007813    -1.5   0.3743
B:C         -0.039062   0.007813    -5.0   0.1257
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0221 on 1 degrees of freedom
Multiple R-squared:  0.9943,    Adjusted R-squared:  0.9604
F-statistic: 29.29 on 6 and 1 DF,  p-value: 0.1405
```

Hide

```
effects2 <- 2*lm2$coeff
print(effects2)
```

```
(Intercept)           A           B           C         A:B         A:C
  0.6015625   0.0234375  -0.0781250  -0.1328125  -0.1093750  -0.0234375
        B:C
 -0.0781250
```

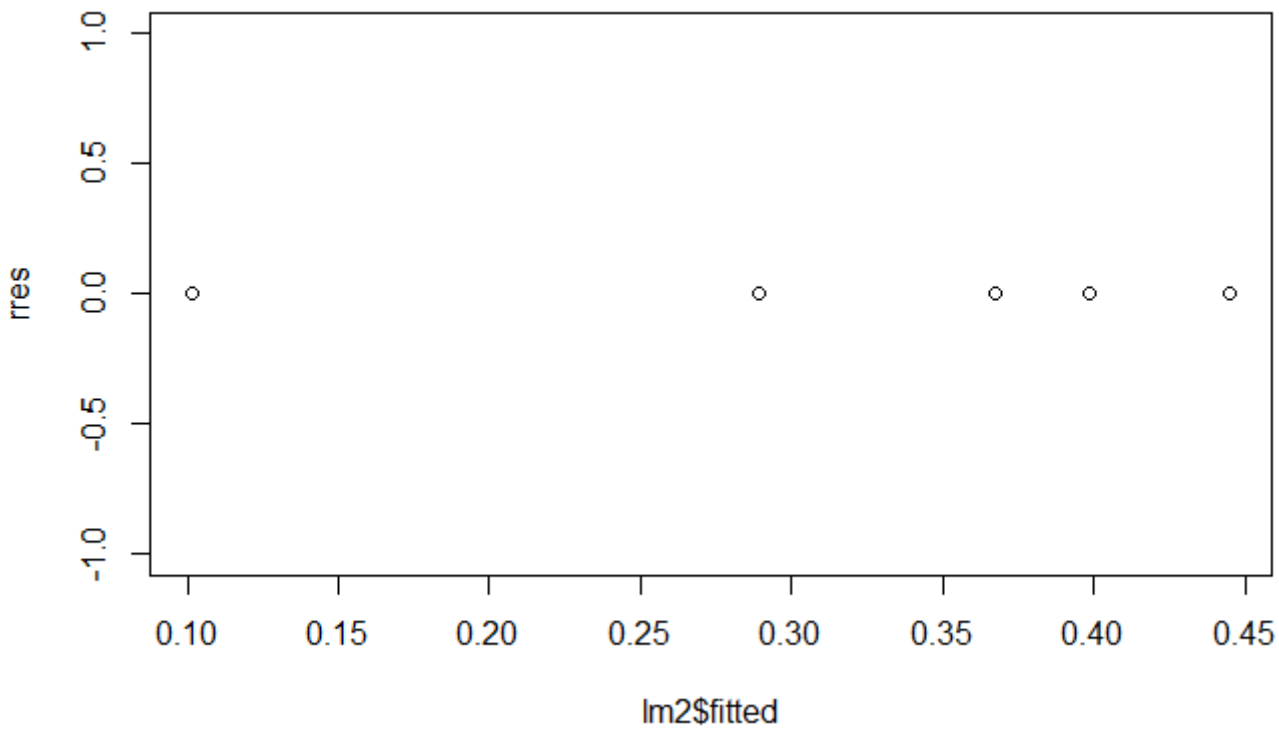
Hide

```
rres <- rstudent(lm2)
print(rres)
```

```
  1   2   3   4   5   6   7   8
  0   0   0 NaN NaN   0 NaN   0
```

Hide

```
plot(lm2$fitted,rres)
```



(I get error in the rstudent() function)

Conclusion and recommendations

I found that all effects and inerations were significant, however we see that the effect of the grade level was large compared to the other level and that a higher grade predicts less spelling mistakes. We also see a low effect from gender like i hypotisized in the begining of the report. I was wrong in my assumtion that there would

be low interactions between the effects. we see that they are all significant. In particular we see that the interaction between gender and age is very high. From the model we se that going from a woman to a man increeses the spelling misstakes, and going from low age to high age decreeses them.