Stats 101B Homework #5

Anna Piskun (505109399) - Discussion 1A

5/22/2020

Question 1

An aluminum master alloy manufacturer produces grain refiners in ingot form. The company produces the product in four furnaces. Each furnace is known to have its own unique operating characteristics, so any experiment run in the foundry that involves more than one furnace will consider furnaces as a nuisance variable. The process engineers suspect that stirring rate affects the grain size of the product. Each furnace can be run at four different stirring rates. A randomized block design is run for a particular refiner, and the resulting grain size data is as follows.

```
data <- read.csv("hw5_Q1.csv")
```

(a) What kind of experimental design is the above experiment? Report your anova table.

The above experiment is randomized block design and follows a fixed effects model (because Stirring Rate and Furnace are fixed).

```
model <- aov(GrainSize~factor(Furnace) + factor(StirringRate), data = data)
summary(model)
```

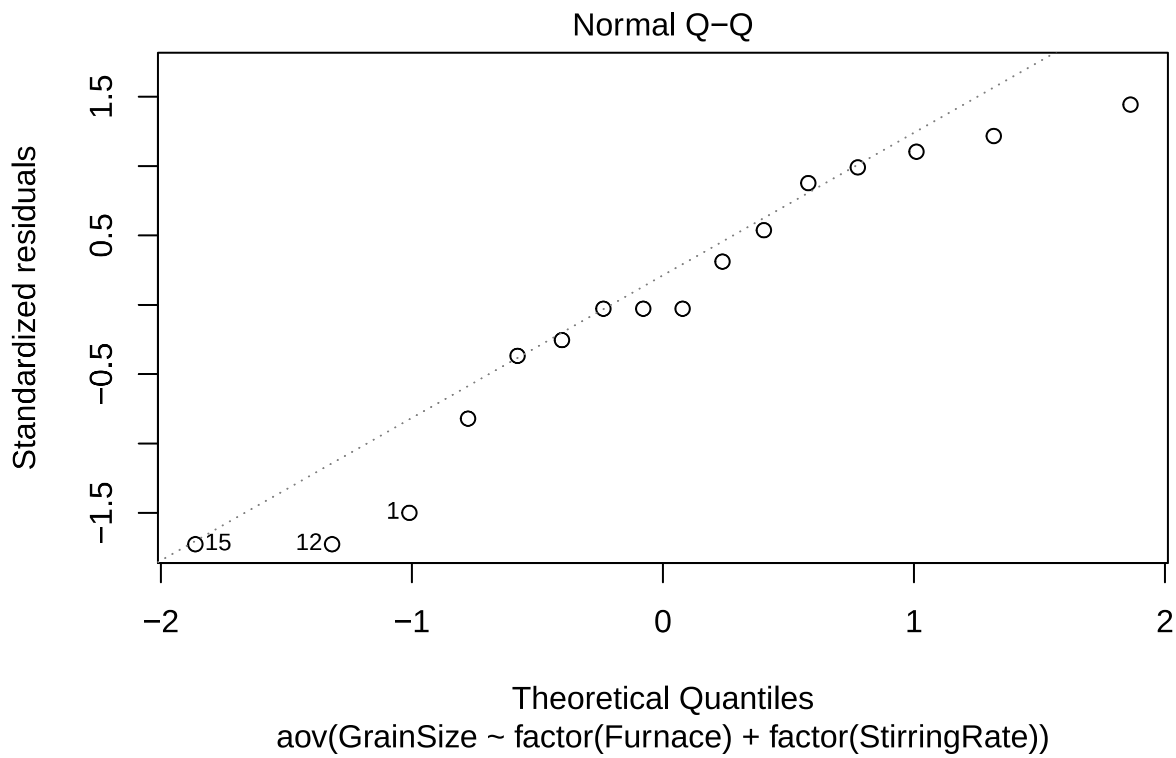
```
##                      Df Sum Sq Mean Sq F value Pr(>F)
## factor(Furnace)       3 165.19   55.06   6.348 0.0133 *
## factor(StirringRate)  3  22.19    7.40   0.853 0.4995
## Residuals             9  78.06    8.67
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) Is there any evidence that stirring rate affects grain size?

Looking at the above anova table, since the pvalue for the stirring rate treatment factor is greater than 0.05, we fail to reject the null hypothesis meaning that there is no difference in grain size as a result of different stirring rates.

(c) Graph the residuals from this experiment on a normal probability plot. Interpret this plot.

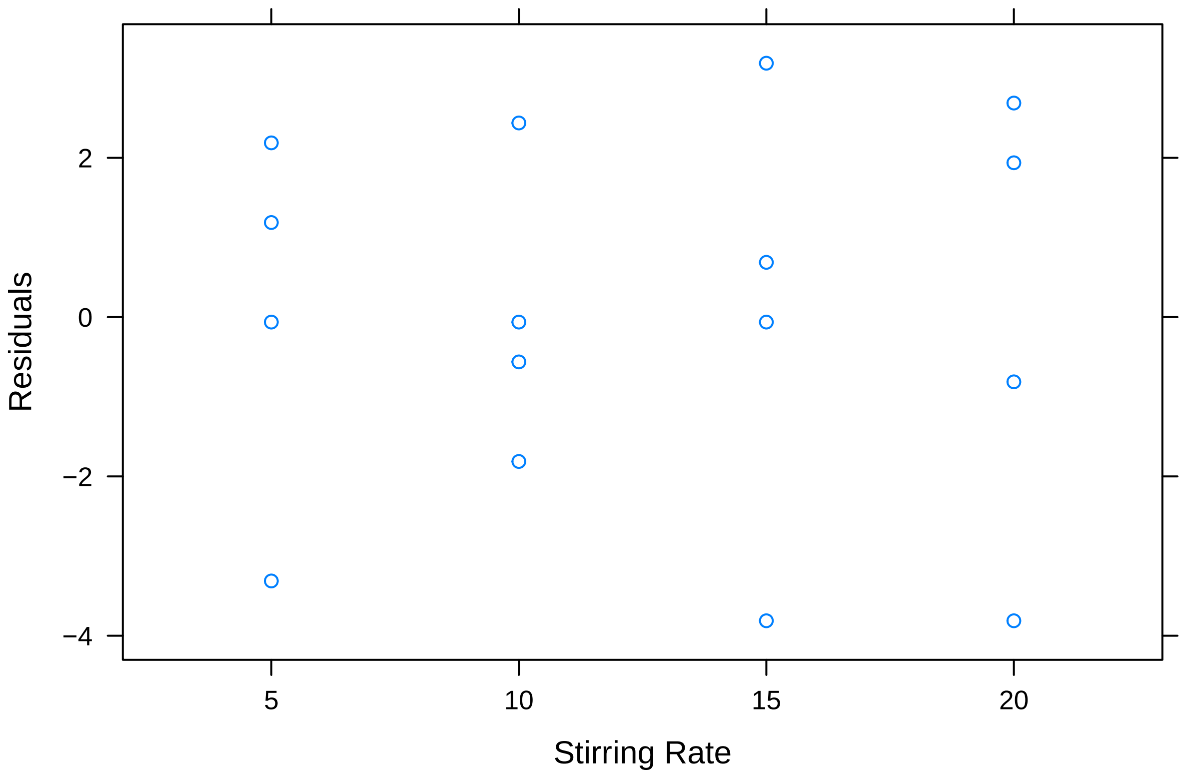
```
plot(model, which = 2)
```



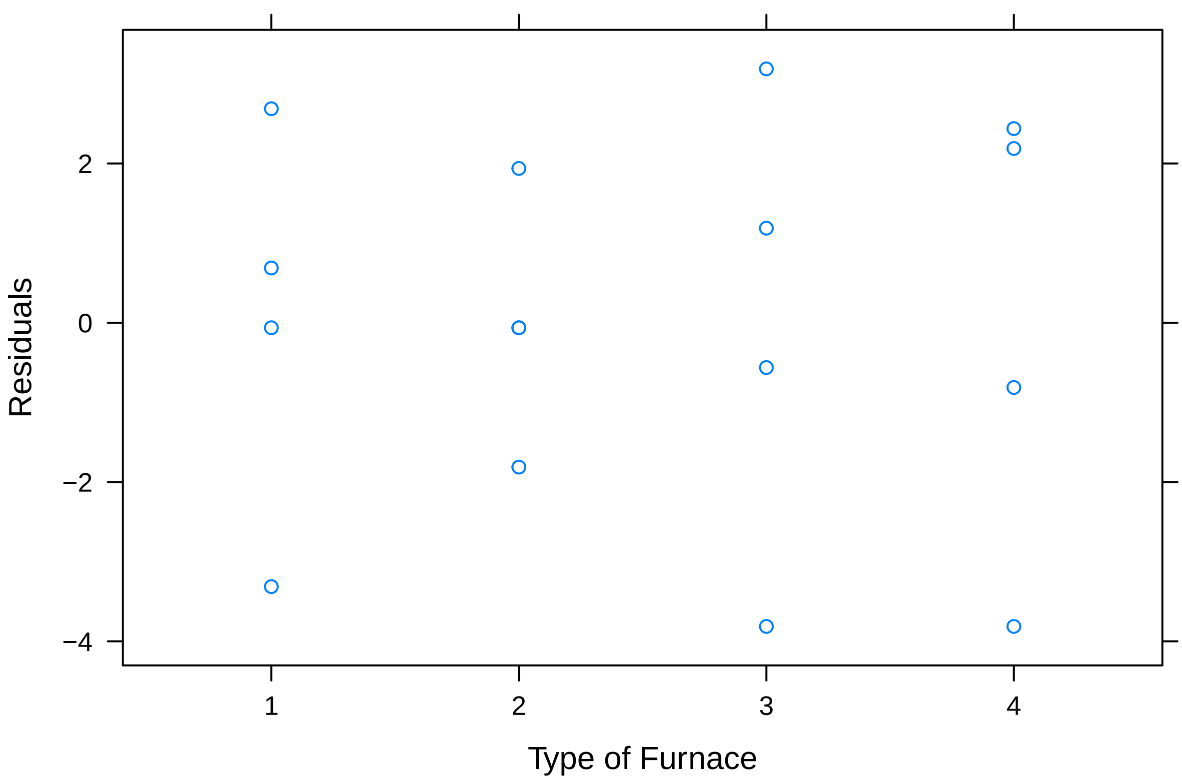
Looking at the normal probability plot above, there is significant deviation from a straight line indicating that the errors are not normally distributed. As such, the normality assumption is not valid.

(d) Plot the residuals versus furnace and stirring rate. Does this plot convey any useful information?

```
library(lattice)
df1 <- data.frame(x = as.factor(data$StirringRate), y = model$residuals)
xyplot(y~x, data = df1,
        xlab = "Stirring Rate",
       ylab = "Residuals")
```




```r
df2 <- data.frame(x = as.factor(data$Furnace), y = model$residuals)
 xyplot(y~x, data = df2,
        xlab = "Type of Furnace",
        ylab = "Residuals")
```



Analyzing the plots for residuals vs. furnace and residuals vs. stirring rate, we see that there are equal points above and below 0 at each stirring rate indicating good model fit. Additionally, the fluctuations of residuals within each factor are similar across all four levels indicating that the constant variance assumption is satisfied.

(e) What should the process engineers recommend concerning the choice of stirring rate and furnace for this particular grain refiner if small grain size is desirable?

```
by(data$GrainSize, factor(data$StirringRate), function(x) mean(x))
```

```
## factor(data$StirringRate): 5
## [1] 5.75
## ------------------------------------------------------------
## factor(data$StirringRate): 10
## [1] 8.5
## ------------------------------------------------------------
## factor(data$StirringRate): 15
## [1] 7.75
## ------------------------------------------------------------
## factor(data$StirringRate): 20
## [1] 8.75
```

```
by(data$GrainSize, factor(data$Furnace), function(x) mean(x))
```

```
## factor(data$Furnace): 1
## [1] 13.25
## ------------------------------------------------------------
## factor(data$Furnace): 2
## [1] 6
## ------------------------------------------------------------
## factor(data$Furnace): 3
## [1] 5.75
## ------------------------------------------------------------
## factor(data$Furnace): 4
## [1] 5.75
```

Looking at the above marginal averages (although stirring rate does not prove to be significant in effecting grain size) engineers should recommend a stirring rate of 5 rpm and either Furnace 3 or 4 (both produce similar grain sizes).

Question 2

The effect of five different ingredients (A, B, C, D, E) on the reaction time of a chemical process is being studied. Each batch of new material is only large enough to permit five runs to be made. Furthermore, each run requires approximately hours, so only five runs can be made in one day. The experimenter decides to run the experiment as a Latin square so that day and batch effects may be systematically controlled. She obtains the data that follow. Analyze the data from this experiment (use ($= 0.05$) and draw conclusions

```
y <- c(8,11,4,6,4,7,2,9,8,2,1,7,10,6,3,7,3,1,6,8,3,8,5,10,8)
batch <- as.factor(rep(c(1:5), times = 5))
day <- as.factor(c(rep(1,5), rep(2,5), rep(3,5), rep(4,5), rep(5,5)))
ingredients <- as.factor(c("A", "C", "B", "D", "E", "B", "E", "A", "C", "D", "D", "A", "C", "E", "B", "
df <- data.frame(y, batch, day, ingredients)
df
```

```
##    y batch day ingredients
## 1  8     1   1           A
## 2 11     2   1           C
```

```
## 3   4     3   1           B
## 4   6     4   1           D
## 5   4     5   1           E
## 6   7     1   2           B
## 7   2     2   2           E
## 8   9     3   2           A
## 9   8     4   2           C
## 10  2     5   2           D
## 11  1     1   3           D
## 12  7     2   3           A
## 13 10     3   3           C
## 14  6     4   3           E
## 15  3     5   3           B
## 16  7     1   4           C
## 17  3     2   4           D
## 18  1     3   4           E
## 19  6     4   4           B
## 20  8     5   4           A
## 21  3     1   5           E
## 22  8     2   5           B
## 23  5     3   5           D
## 24 10     4   5           A
## 25  8     5   5           C
```

```
model.df <- lm(y ~ factor(batch) + factor(day) + factor(ingredients), data = df)
anova(model.df)
```

```
## Analysis of Variance Table
##
## Response: y
##                     Df Sum Sq Mean Sq F value    Pr(>F)
## factor(batch)        4  15.44   3.860  1.2345 0.3476182
## factor(day)          4  12.24   3.060  0.9787 0.4550143
## factor(ingredients)  4 141.44  35.360 11.3092 0.0004877 ***
## Residuals           12  37.52   3.127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model.av <- aov(model.df)

treatment.means <- vector(length = 5)

for(j in c("A", "B", "C", "D", "E")) {
  treatment.means[j] <- mean(y[ingredients==j])
}

treatment.means
```

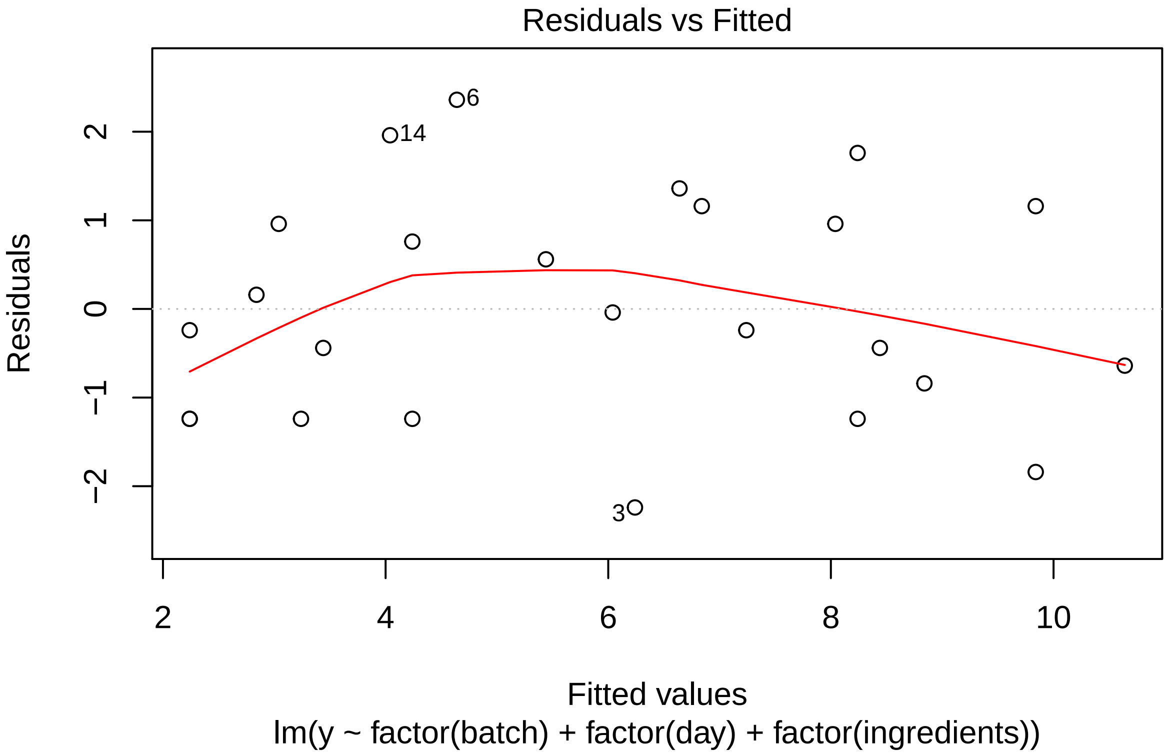
```
##                       A   B   C   D   E
## 0.0 0.0 0.0 0.0 0.0 8.4 5.6 8.8 3.4 3.2
```

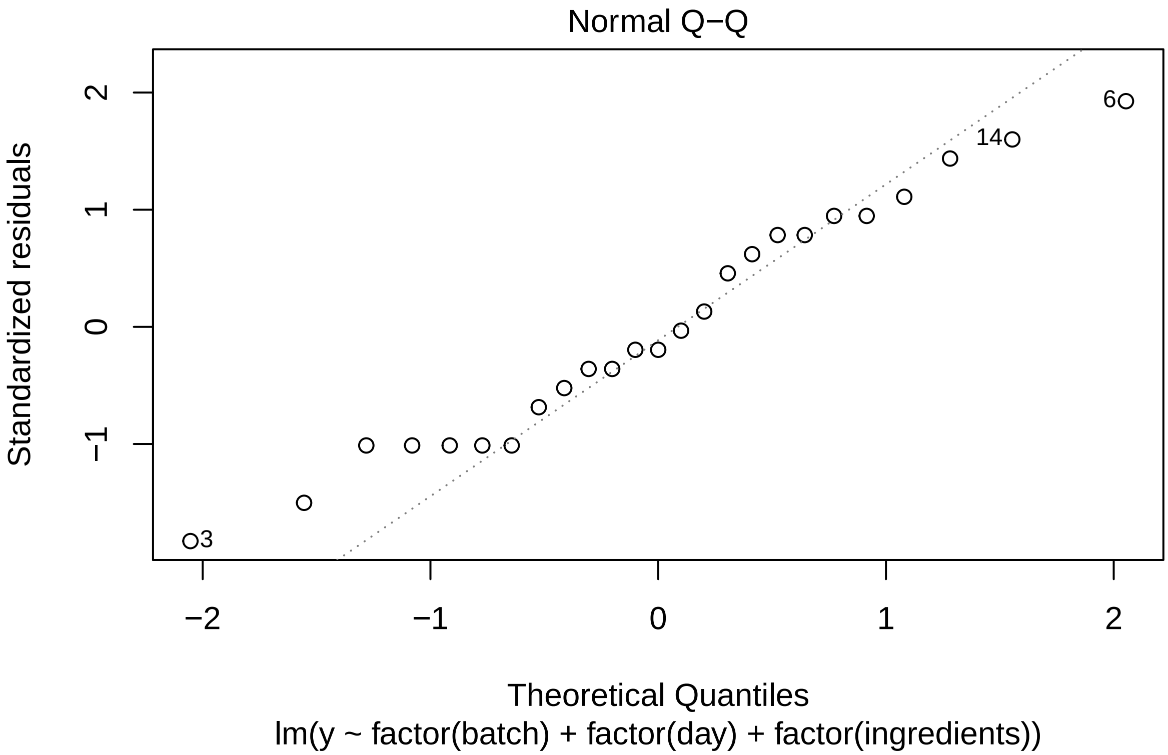
```
TukeyHSD(model.av)
```

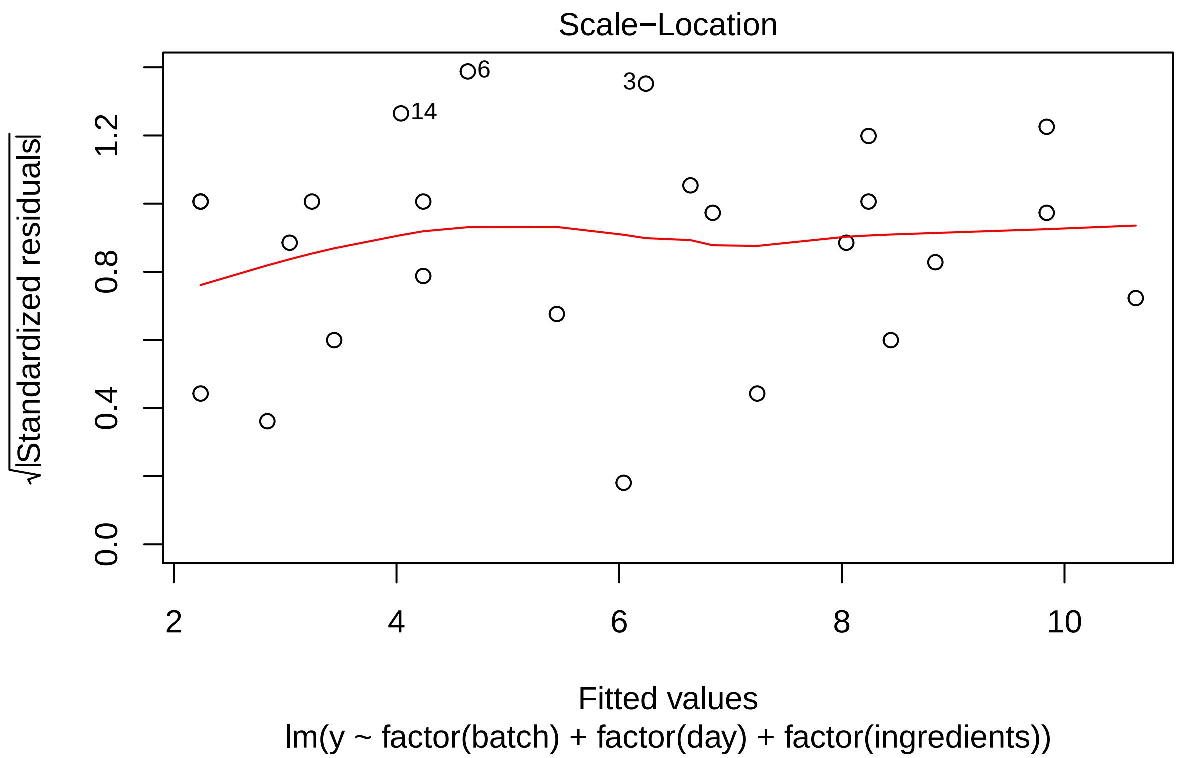
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
```

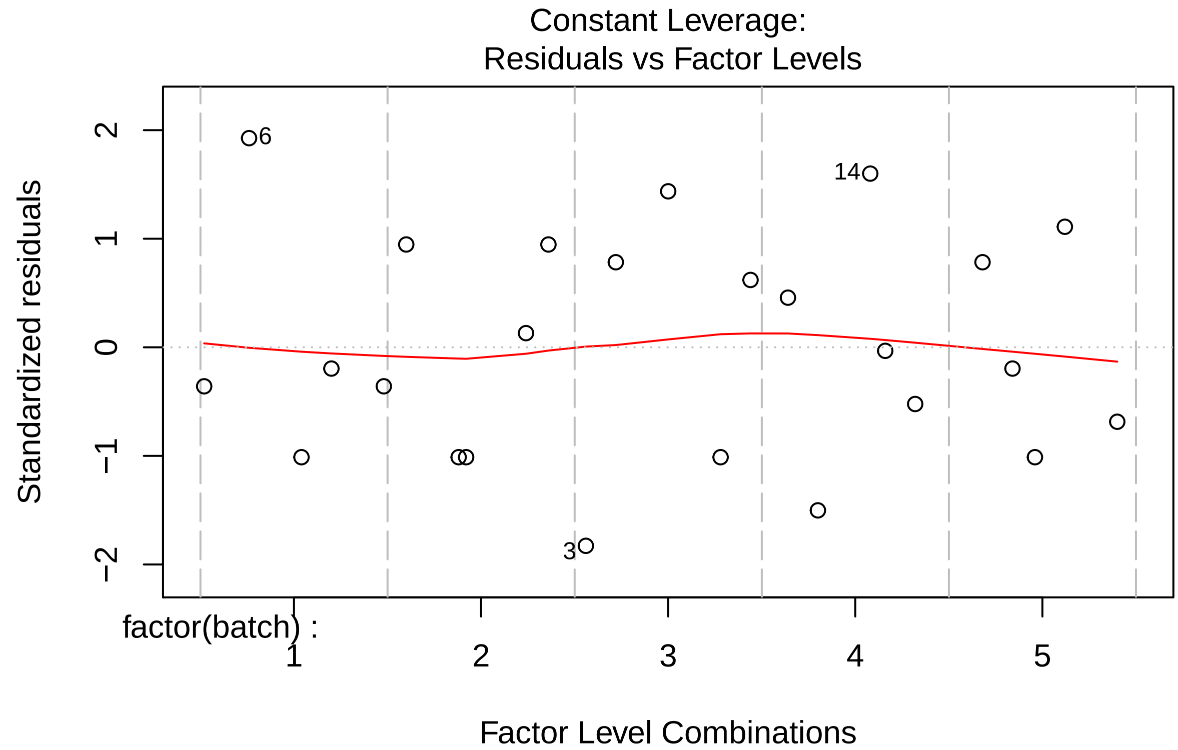
```
## Fit: aov(formula = model.df)
##
## $`factor(batch)`
##     diff       lwr      upr     p adj
## 2-1  1.0 -2.564608 4.564608 0.8936609
## 3-1  0.6 -2.964608 4.164608 0.9816047
## 4-1  2.0 -1.564608 5.564608 0.4225127
## 5-1 -0.2 -3.764608 3.364608 0.9997349
## 3-2 -0.4 -3.964608 3.164608 0.9960012
## 4-2  1.0 -2.564608 4.564608 0.8936609
## 5-2 -1.2 -4.764608 2.364608 0.8166339
## 4-3  1.4 -2.164608 4.964608 0.7232162
## 5-3 -0.8 -4.364608 2.764608 0.9489243
## 5-4 -2.2 -5.764608 1.364608 0.3365811
##
## $`factor(day)`
##     diff       lwr      upr     p adj
## 2-1 -1.0 -4.564608 2.564608 0.8936609
## 3-1 -1.2 -4.764608 2.364608 0.8166339
## 4-1 -1.6 -5.164608 1.964608 0.6212723
## 5-1  0.2 -3.364608 3.764608 0.9997349
## 3-2 -0.2 -3.764608 3.364608 0.9997349
## 4-2 -0.6 -4.164608 2.964608 0.9816047
## 5-2  1.2 -2.364608 4.764608 0.8166339
## 4-3 -0.4 -3.964608 3.164608 0.9960012
## 5-3  1.4 -2.164608 4.964608 0.7232162
## 5-4  1.8 -1.764608 5.364608 0.5188508
##
## $`factor(ingredients)`
##     diff        lwr        upr     p adj
## B-A -2.8 -6.3646078  0.7646078 0.1539433
## C-A  0.4 -3.1646078  3.9646078 0.9960012
## D-A -5.0 -8.5646078 -1.4353922 0.0055862
## E-A -5.2 -8.7646078 -1.6353922 0.0041431
## C-B  3.2 -0.3646078  6.7646078 0.0864353
## D-B -2.2 -5.7646078  1.3646078 0.3365811
## E-B -2.4 -5.9646078  1.1646078 0.2631551
## D-C -5.4 -8.9646078 -1.8353922 0.0030822
## E-C -5.6 -9.1646078 -2.0353922 0.0023007
## E-D -0.2 -3.7646078  3.3646078 0.9997349
```

```r
plot(model.df)
```


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ factor(batch) + factor(day) + factor(ingredients))
6
14
3
2
1
0
−1
−2
2
4
6
8
10


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ factor(batch) + factor(day) + factor(ingredients))
6
14
3
2
1
0
−1
−2
−1
0
1
2





Looking at the above anova table, we see that the only statistically significant factor is ingredients indicating that the mean reaction time as a result of type of chemical ingredient are significantly different. We also can conclude that the effects of batch and day do not have an effect on reaction time since they have pvalues greater than 0.05 meaning we fail to reject the null hypothesis that there is a difference in the average reaction time. Analyzing the output of our Tukey's test, we see that the only significant between group difference is for Ingredients D and A, E and A, D and C, and and E and C.

The residual plot shows no clear pattern or fanshape indicating that the constant variance assumption is satisfied. This is further supported by the scale location plot which shows no increasing or decreasing trend. Likewise, the normal probability plot is fairly straight indicating that the errors are normally distributed and that the normality assumption is satisfied.

Question 3

The yield of a chemical process was measured using five batches of raw material, five acid concentrations, five standing times (A,B,C,D, E) and five catalyst concentrations. The Graeco-Latin square that follows was used. Analyze the data from this experiment (use a = 0.05) and draw conclusions.

```r
data2 <- read.csv("HW5 Q3 data .csv")

data2.av <- aov(Yield~factor(Batch) + factor(Acid) + factor(Time) + factor(Catalyst), data = data2)

summary(data2.av)
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## factor(Batch)     4   10.0    2.50   0.427 0.785447
## factor(Acid)      4   24.4    6.10   1.043 0.442543
## factor(Time)      4  342.8   85.70  14.650 0.000941 ***
## factor(Catalyst)  4   12.0    3.00   0.513 0.728900
## Residuals         8   46.8    5.85
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(data2.av)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Yield ~ factor(Batch) + factor(Acid) + factor(Time) + factor(Catalyst), data = da
##
## $`factor(Batch)`
##     diff       lwr      upr     p adj
## 2-1 -0.2 -5.484751 5.084751 0.9999182
## 3-1 -0.8 -6.084751 4.484751 0.9823986
## 4-1 -1.4 -6.684751 3.884751 0.8834072
## 5-1 -1.6 -6.884751 3.684751 0.8279246
## 3-2 -0.6 -5.884751 4.684751 0.9939694
## 4-2 -1.2 -6.484751 4.084751 0.9281909
## 5-2 -1.4 -6.684751 3.884751 0.8834072
## 4-3 -0.6 -5.884751 4.684751 0.9939694
## 5-3 -0.8 -6.084751 4.484751 0.9823986
## 5-4 -0.2 -5.484751 5.084751 0.9999182
##
## $`factor(Acid)`
##     diff       lwr      upr     p adj
## 2-1 -0.2 -5.484751 5.084751 0.9999182
## 3-1  0.6 -4.684751 5.884751 0.9939694
## 4-1 -1.2 -6.484751 4.084751 0.9281909
## 5-1 -2.2 -7.484751 3.084751 0.6232282
## 3-2  0.8 -4.484751 6.084751 0.9823986
## 4-2 -1.0 -6.284751 4.284751 0.9610846
## 5-2 -2.0 -7.284751 3.284751 0.6948188
## 4-3 -1.8 -7.084751 3.484751 0.7640759
```

```
## 5-3 -2.8 -8.084751 2.484751 0.4197369
## 5-4 -1.0 -6.284751 4.284751 0.9610846
##
## $`factor(Time)`
##      diff        lwr        upr     p adj
## B-A  -8.0 -13.284751 -2.7152488 0.0051639
## C-A  -4.8 -10.084751  0.4847512 0.0770797
## D-A  -8.6 -13.884751 -3.3152488 0.0032815
## E-A -10.6 -15.884751 -5.3152488 0.0008219
## C-B   3.2  -2.084751  8.4847512 0.3087034
## D-B  -0.6  -5.884751  4.6847512 0.9939694
## E-B  -2.6  -7.884751  2.6847512 0.4837165
## D-C  -3.8  -9.084751  1.4847512 0.1869031
## E-C  -5.8 -11.084751 -0.5152488 0.0317351
## E-D  -2.0  -7.284751  3.2847512 0.6948188
##
## $`factor(Catalyst)`
##     diff       lwr      upr     p adj
## b-a  0.4 -4.884751 5.684751 0.9987373
## c-a  1.6 -3.684751 6.884751 0.8279246
## d-a -0.2 -5.484751 5.084751 0.9999182
## e-a  1.2 -4.084751 6.484751 0.9281909
## c-b  1.2 -4.084751 6.484751 0.9281909
## d-b -0.6 -5.884751 4.684751 0.9939694
## e-b  0.8 -4.484751 6.084751 0.9823986
## d-c -1.8 -7.084751 3.484751 0.7640759
## e-c -0.4 -5.684751 4.884751 0.9987373
## e-d  1.4 -3.884751 6.684751 0.8834072
```

Looking at the above anova table we see that the only statistically significant factor is Time. Since its pvalue (essentially 0) is less than 0.05 we reject the null hypothesis and conclude that standing time effects the mean yield of a chemical process. Batch, Acid, and Catalyst, however, all have pvalue greater than 0.05 meaning we fail to reject the null hypothesis and they do not have an effect on the mean yield of a chemical process. The only significant between group differences is for Time B and Time A, Time D and A, Time E and C, and Time E and A.

Question 4

Many people now read on materials provided in an electronic format. Are Kindles, tablets, and computers, just as good as books? Both formats have their advantages and disadvantages, but suppose we want to know the impact on reading retention. We wish to know if people can recall what they've read equally well in all four formats, or whether there are differences in recall. A secondary issue is whether recall is affected by the content of the reading. For this reason, we will prepare two excerpts for the subjects to read. One excerpt will be from a work of action, and the other from the news. Both excerpts are the same length (in terms of numbers of lines of text.)

As "units", we have n children at a high school to serve as subjects (roughly ages ranging from 13 to 18). Students whose parents give them permission will be pulled out of class to participate in the study. We have an instrument (the technical term for an exam that measures some characteristic of a person) that measures how much of each reading sample the child remembers. High scores correspond to high memory retention.

You will be asked to design three studies, and to compare them. You are welcome to make any assumptions you wish, but need to state them explicitly. For example, you might want to assume that a certain quality can be measured and/or observed.

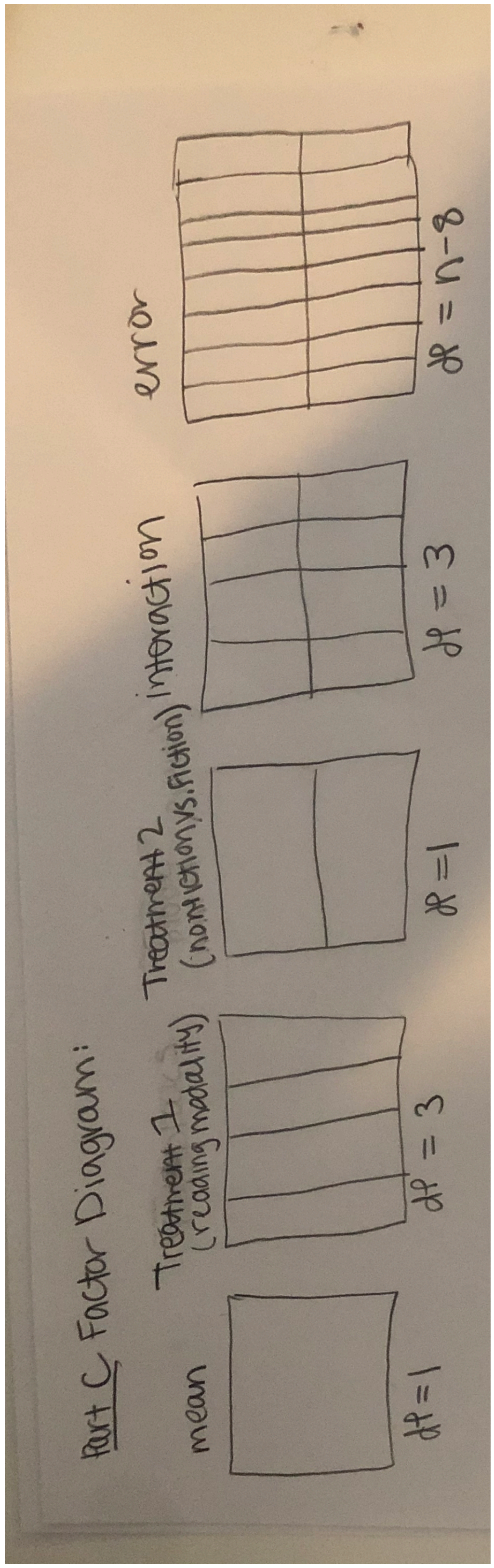
A. The description above mentions (at least) two nuisance factors, some of which we've addressed. What are they and why do you think they might be potential confounding factors?

One nuissance factor is age. If the subjects are ranging from 13-18 that can result in varying comprehension levels due to age and academic rigor in class. For example, an 18 year old would be around a senior in high school about to go to college while a 13 year old would be just starting high school. This age difference would definitely serve as a confounding factor. Another nuissance factor is the length of the passage as some lengths might be better suited for different mediums. Likewise, longer passages may be harder to focus on and as a result students will retain less information than that compared to a shorter passage.

B. Name two other possible nuisance factors

Another possible nuisance factor is the subject area of material. For example, if you love science and read an article about science, you are more likely to retain more of information from it than someone who loves history and hates science. A fourth possible nuisance factor is the time of day at which the reading material is administered. For example, it might be harder to retain information later in the day as students may be tired vs. first thing in the morning when students are more fresh and alert.

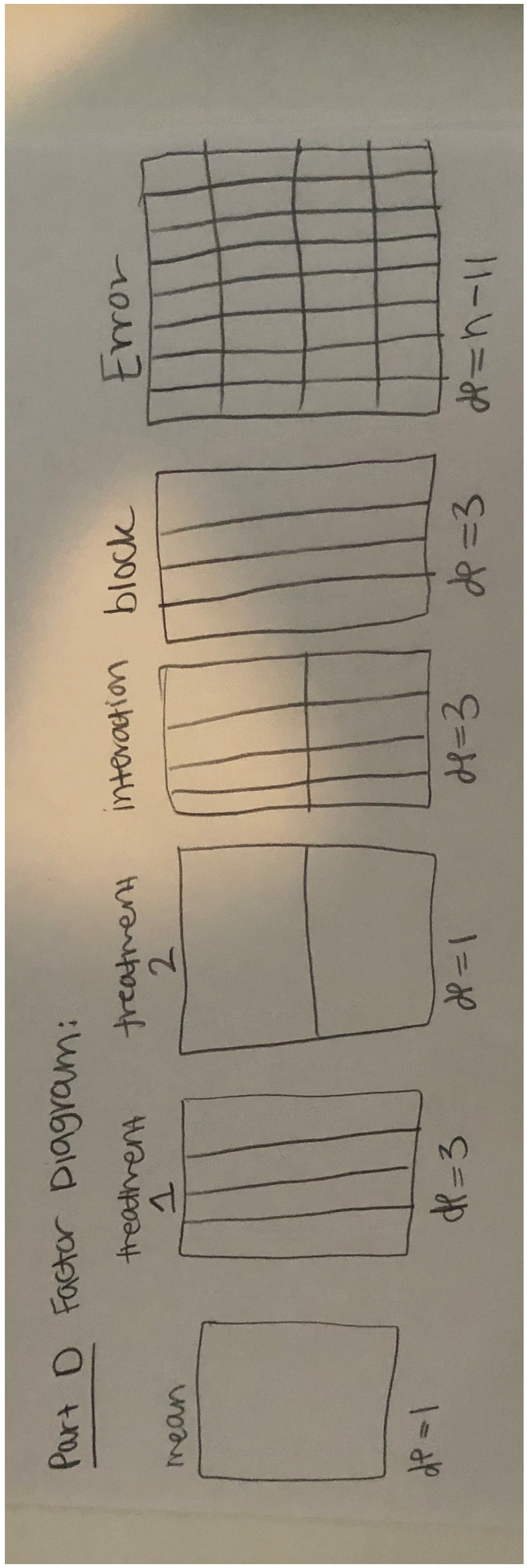
C. Describe a two-way completely randomized design (with no blocking)



Since the degrees of freedom for the errors is n-8, we need to have more than 8 students or else the residuals will have 0 degrees of freedom and we can't quantify the errors. As such, we need a multiple of 8 based off of the interaction. Here I chose 16 because its the smallest number of students that allows for a balanced design with 2 students per treatment.

To carry out randomization, I would assign the subjects a number 1-16. The numbers are written on cards, shuffled, then dealt into four piles A (treatment 1-kindles), B (treatment 2-tablets), C (treatment 3-computers), and D (treatment 4-books) respectively. Then I would assign the piles a number 1 through 4, write the numbers on cards, shuffle them, and then seperate into two piles. Pile 1 corresponds to Fiction and Pile 2 corresponds to Non-Fiction. Utilizing r code I would write: Sample(1:16,16).

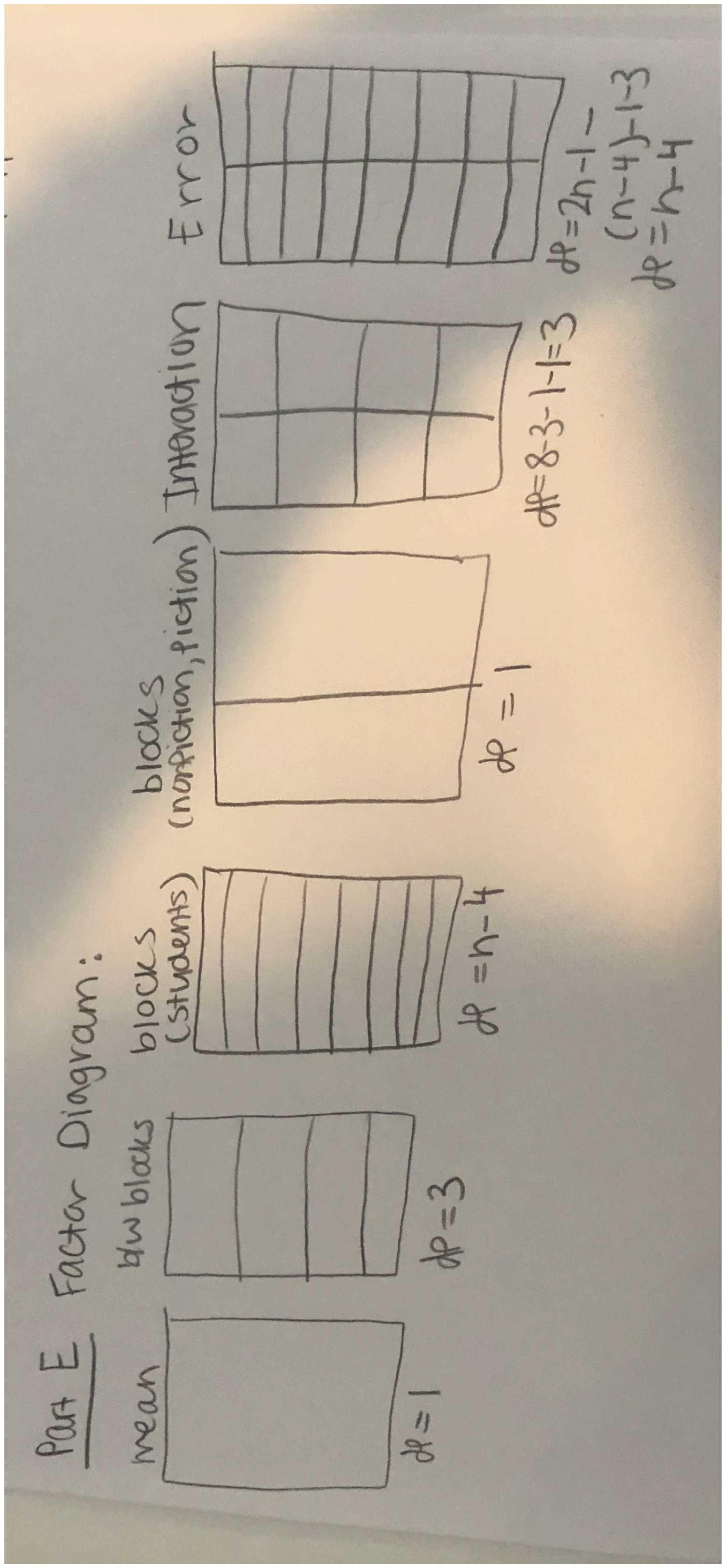
D. Describe a two-way randomized completely block design (for one nuisance factor)



The blocking factor in this design would be by grade. I would have four blocks: freshman, sophomore, junior, and senior. I believe this will make a good blocking factor as a nuisance factor discussed above was the reading comprehension level variability for students in different grades and of different ages (a senior in high school vs. a freshman in high school). Through blocking by grade level, I will control for some of the outside variance that may affect reading comprehension.

I would first separate by block so there would be four groups: freshman, sophomore, junior, and senior each with 8 students. I would then assign each student a number 1-8, write them down on cards, shuffle, then deal equally into eight piles (which correspond to their interaction combination). Lastly, I'll assign each pile a number 1-8, write it down on cards, shuffle, and then seperate into another four piles. Each pile corresponds to a reading treatment. Then I would assign each pile a number 1-4, write it down on cards, shuffle, and then seperate into 2 piles. Pile 1 correponds to fiction and Pile 2 corresponds to Nonfiction. Utilizing r code I would write Sample(1:8, 8) and repeat to randomly sample within each block.

E. Describe a split plot/repeated measures design



Using split plot/repeated measures design you measure the same person twice for two different treatments (total number of data points = 32) - every student is a block, but whole purpose of switching to this design is because researchers suspect person to person variability is high (rather than doing latin squares).

Therefore, first I would randomly assign the outside factor which in this case would be reading format. Using sample(1:8,8) the first two numbers get Kindle, the second 2 get Tablet, the third 2 get computer, and last 2 get book. Then, I would randomize the inside factor which in this study would be fiction vs. non fiction. To do so, I flip a coin for each student with heads being reading fiction first (non-fiction second) and tails being reading non-fiction first (fiction second).

F. Which study would you recommend and why?

I would recommend split plot/repeated measures design as it allows researchers to see how one specific test subject changes as a result of the treatment factors and controls more for person to person variability which is high in this type of study.