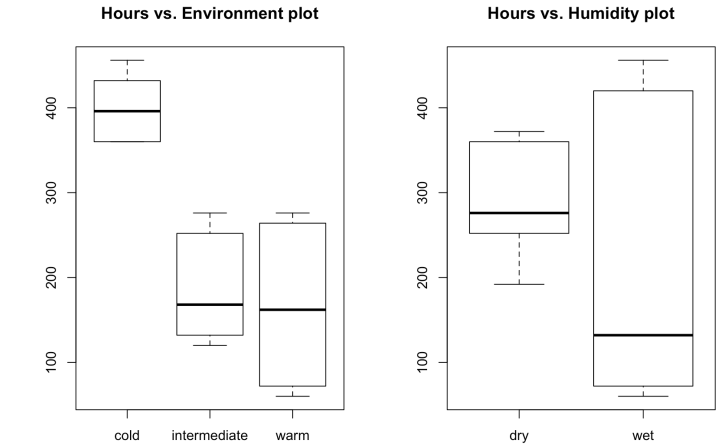
Assignment 3

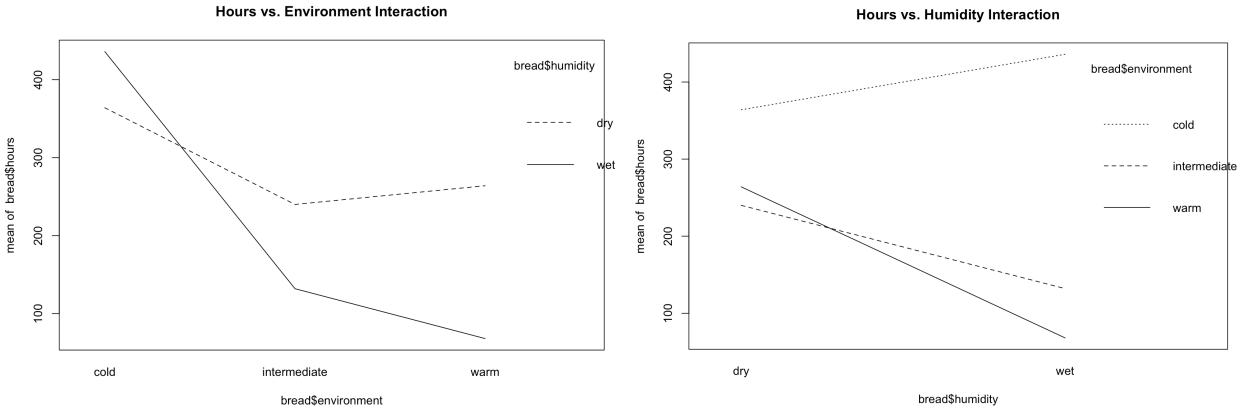
Chao Zhang & Ibrahim Kanj
Group 13

Exercise 1)

1) Check appendix (R-code)

2)

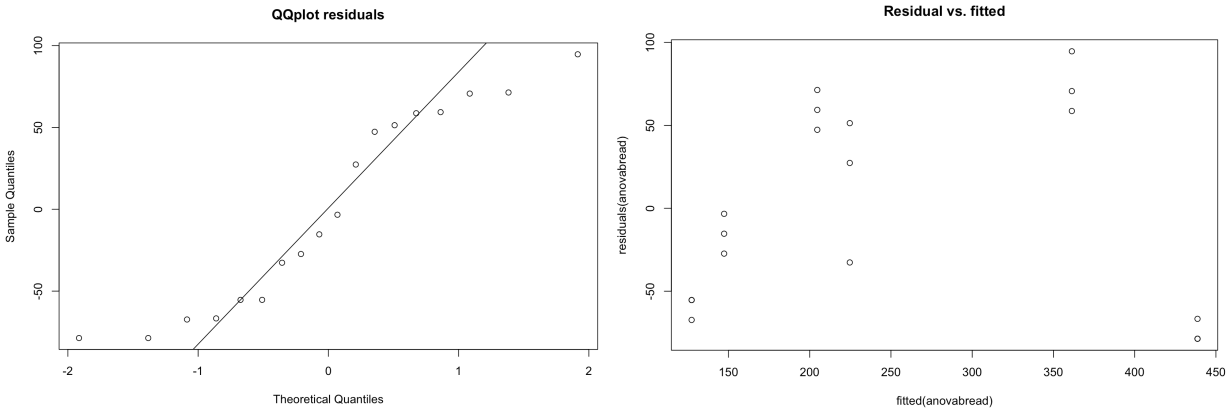




3) By looking at the model for environment and humidity alone, we see that the environment plays a bigger effect on the bread than the humidity (p-value for environment: 3.674e-05, p-value for humidity: 0.02637). Now if we look at the interaction model between the environment and humidity, we can notice that the interaction between environment and humidity has higher significance and effect on bread than the action of humidity alone.
By looking at the interaction plot the significance of interaction becomes clear where we see the huge difference between cold/wet and warm/wet.

4) By looking at the summary of anova we can see three effects, the effect of temperature (intermediate,warm) and humidity (wet). we can say that temperature's effect is more vital on the decay of bread. This question is important because if one factor has a bigger effect on the decay of bread than the other, then it's better to deal with this factor first.

5)

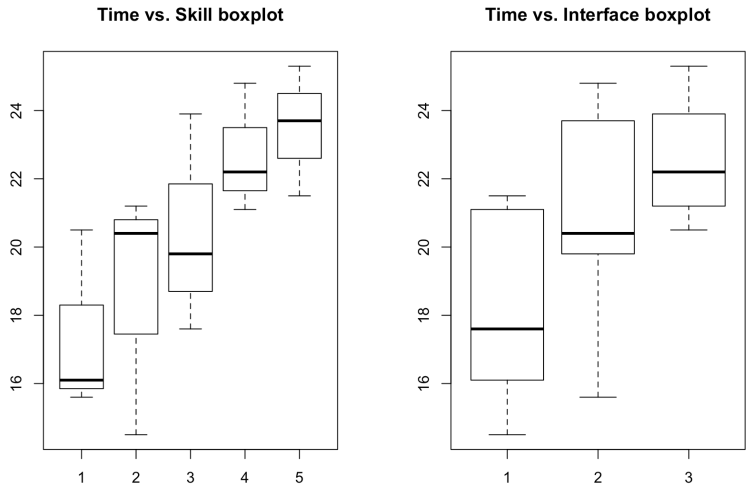


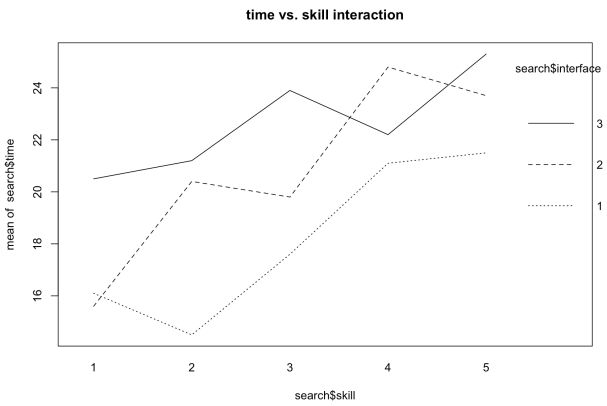
The residual's plot seems somehow normal by looking at the qqplot. By looking at the residual vs. fitted plot there seems to be some outliers where there are points going above 70.

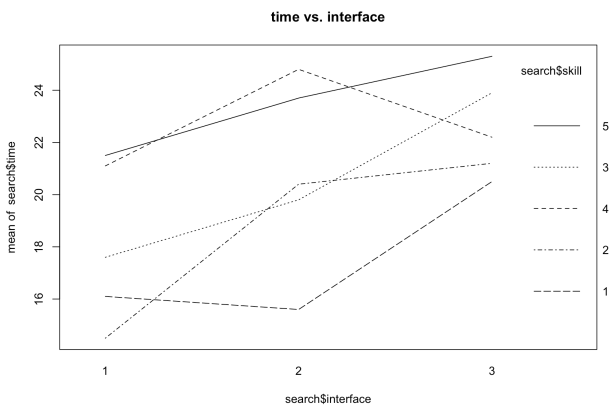
Exercise 2)

1) Check appendix (R-code)

2)





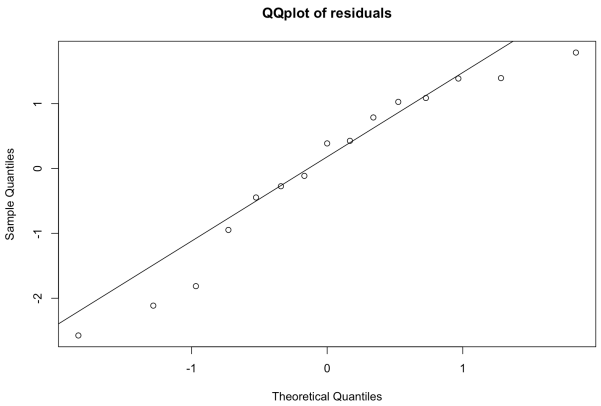


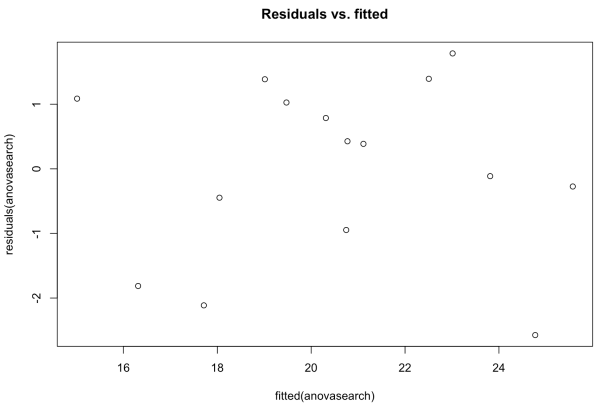
By looking at the interaction plot we could see that users with good computer handling managed to have lower times in all three interfaces when compared with the other users of different skills. Another observation is that interface 3 seems to be the hardest for users to understand, where users of different skills took about 20 - 25 seconds.

3) The null hypothesis is rejected where p-value for the interface was 0.01310. This means that the interfaces plays a role in the time user spend using the search engine.

4) By looking at the summary of the anova we can see that user 4 spends 5.3 seconds more than a user with skill 1. interface 3 requires 4.46 more time than interface 1.

5)





By looking at the plot, the residuals seems normal and thus the anova is valid.

6) The p-value = 0.04076 < 0.05 implies the null hypothesis can be rejected meaning that the interface has an effect on the time user spends on the search engine.

7) The p-value (0.09642) is larger than 0.05 so we could not reject the null hypothesis that says that the search time is the same for all interfaces. the residuals are normally distributed so the anova is valid. However, an assumption of doing one way anova is to have a completely randomized design, but we know that this data come from users that are organized according to

5 different experience levels so trying one way anova doesn't really add anything here. But if we repeat it and we choose random 15 users then we might answer whether the differences in interfaces affects the search time.

Exercise 3)

1) By looking at the p-values we can say that both starter and batch have an effect on the formation of bacteria but starter has even higher significance than batch where the p-values were 2.904e-05 and 0.001632 respectively.

2) By looking at the summary of anova we can see that starters 1 & 4 have p-values less than 0.05 and thus they have an effect in the formation of bacteria.

4)

```
 > confint(anovacream)
                                2.5 %       97.5 %
(Intercept)                 7.5006177   9.82258232
factor(cream$starter)2     -1.1682489   0.86824885
factor(cream$starter)3     -1.9982489   0.03824885
factor(cream$starter)4      1.7917511   3.82824885
factor(cream$starter)5     -1.5022489   0.53424885
```

starter 4 doesn't include 0 and that's in accordance with the results back in #2 where actually starter 4 causes the most formation of bacteria, thus it show values higher than that of starter 1, while the others (starter 2,3 and 5) are all in between starter 1.

Exercise 4)

1) By looking at both anova(cowlm) and summary(cowlm) we see that p-value is not less than 0.05 and thus treatment has no significant effect on the production of milk.

2) We see that by using treatment B we actually get less milk than treatment A by 0.51
also period 2 leads to less milk production by 2.39
as for the order BA we also see less milk production than AB by 11.2

3)

- by looking at anova(cowlmer1,cowlmer) we see that the p-value is greater than 0.05 then treatment has no significant effect on milk production similar to the fixed effects.
- the mixed effect has similar estimation as #2 for treatment B and order BA however for period 2 we see less milk produced by 3.47

4) The paired t-test would provide meaningful results if the neutral period allows for truly washing out the carry-over effect of taking a certain treatment before the other. This test could

determine whether there is a significant difference between the treatments on the production of milk.
The resulting p-value (0.8281) of the t-test is greater than 0.05 and thus there is no significance difference on the production of milk whether using treatment A or B. this is in line with the results we found in #1.

Exercise 5)

1) Check appendix (R-code)

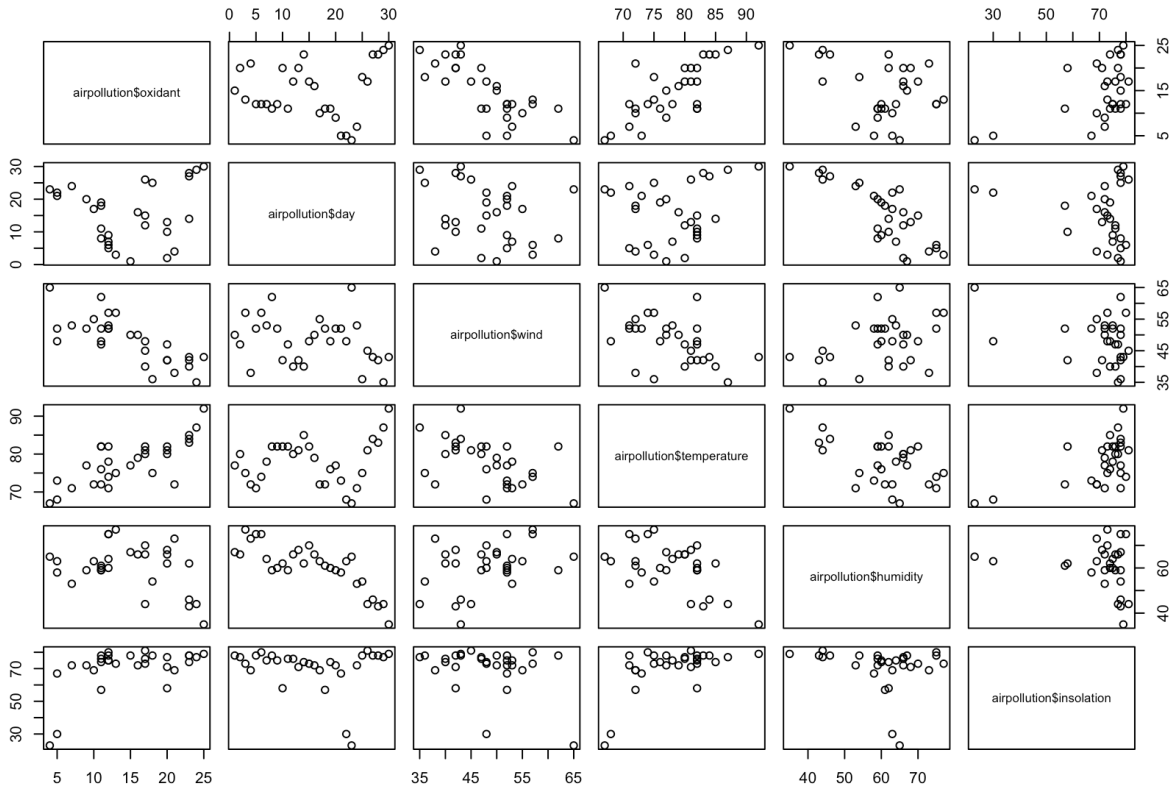
2) The outcome of xtabs(~medicine,nausea) is similar to text file that we're given nauseatable.txt

3) By applying anova on the data we see a p-value = 0.2571 for the medicine so there is no significance effect for medicine on nausea

4) By looking at the p-value (0.03643) of the chi-square we can say that there is a significance effect for the medicine on inducing nausea.

Exercise 6)

1)



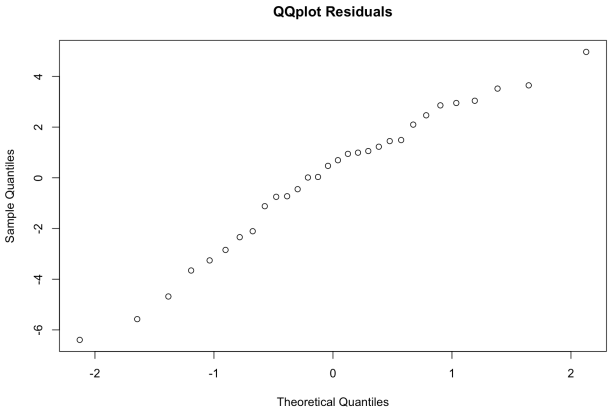
By looking at the different plots and the slopes, it looks like there is some big correlation between (oxidation and wind), (oxidation and temperature),(oxidation and isolation), (day and humidity), (temperature and wind),(temperature and humidity) and in these specific plots the points are close to one another (no real outliers) so a regression model could prove useful to explain the different interactions.

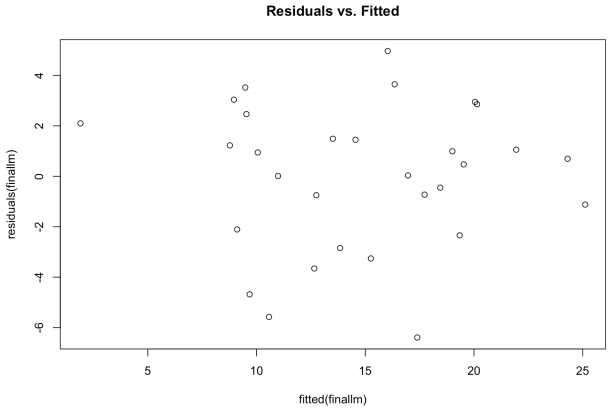
2) By looking at the p-values of the added explanatory variables in step 3 (insolation, humidity and day) they are all greater than 0.05 so we had to stop and go back to the best result from step 2. thus the linear regression model is :
lm(airpollution$oxidant~airpollution$wind+airpollution$temperature,data=airpollution)

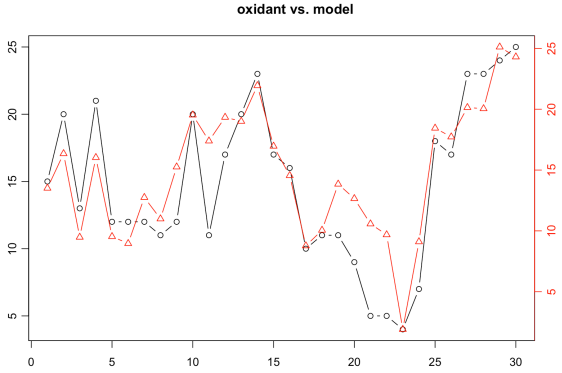
3) The linear regression model that is dependent on all explanatory variables would be:
-12.0401 - 0.44749*wind + 0.55714*temperature + 0.01822*insolation - 0.02997*day + 0.06818 but of course this is not the best formula because there are explanatory variables with p-value greater than 0.
The highest p-value is that of day so we remove day. Next we remove insolation. then we remove humidity. And we're left with the same linear regression model as #2
lm(airpollution$oxidant~airpollution$wind+airpollution$temperature,data=airpollution)

4) oxidation = -5.203445 - 0.42706*wind + 0.52035*temperature + error

5)

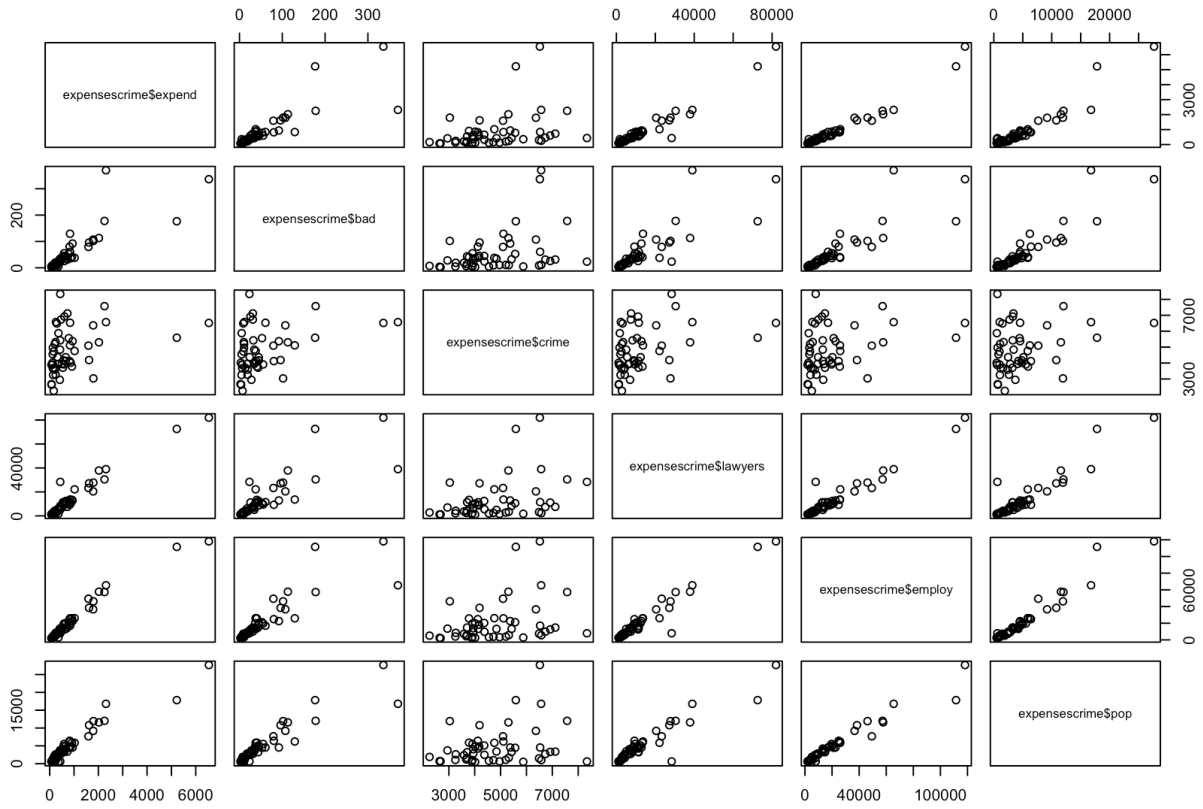






According to the plots the residuals of the chosen model are normal and so It's appropriate to use the model as an estimation for the oxidation and there's in an error between the range [-7,5] as seen from the second plot.

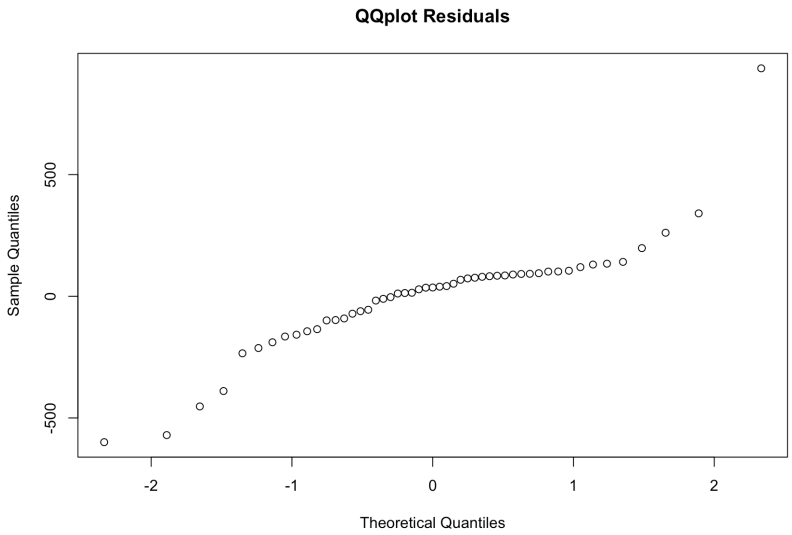
Exercise 7)



By looking at the graphs pop, lawyers, employ appear to be very collinear . By looking at the numerical values we see indeed large values (pop-lawyers : 0.93, employ-lawyers: 0.97, pop-employ: 0.97) even numerical values show bad-pop: 0.92 which is also high. Thus the model will have to drop these variables.

We've decided to do step-up method to find the linear regression model.
the linear model is -0.01107 + 0.02971*employ + 0.02686*lawyers

By looking at cook's distance we notice two influence points. point 5 and point 8. when we look at the residuals of these points we see 936.54642 and -452.66652 respectively which is really big.



If we look at the plot for the residuals of the fitted model, we notice that the graph is not normal. Removing the influence points doesn't fix the normality of the model. Thus the model can not estimate the expenditure on criminal activities.

APPENDIX

Full R Code

```
remove(list = ls())
par(mfrow=c(1,1))

install.packages('multcomp')
install.packages('lme4')
library(multcomp)
library(lme4)

#Exercise 1)
bread = read.table("bread.txt",header=TRUE)

#1)
environment = c("cold","intermediate","warm")
humidity = c("dry","wet")
N = 18
cases = N / (length(environment)*length(humidity))
for (i in 1:length(environment))
{
  for (b in 1:length(humidity) )
  {
    for (c in 1:cases)
    {
      cat(environment[i],humidity[b],sep = " ",fill = TRUE)
    }
  }
}

#2)
xtabs(bread$hours~bread$environment+bread$humidity,data=bread)
par(mfrow=c(1,2))
boxplot(bread$hours~bread$environment) + title("Hours vs. Environment plot")
boxplot(bread$hours~bread$humidity) + title("Hours vs. Humidity plot")

par(mfrow=c(1,1))
interaction.plot(bread$environment,bread$humidity,bread$hours) + title("Hours vs. Environment
Interaction")
interaction.plot(bread$humidity,bread$environment,bread$hours) + title("Hours vs. Humidity
Interaction")
```

#3)

```
anovabread=lm(bread$hours~bread$environment+bread$humidity,data=bread)
anova(anovabread)

anovabreadinteraction = lm(bread$hours~bread$environment*bread$humidity,data=bread)
anova(anovabreadinteraction)
```

#4)

```
summary(anovabread)
```

#5)

```
qqnorm(residuals(anovabread),main ="QQplot residuals")
qqline(residuals(anovabread))
plot(fitted(anovabread),residuals(anovabread),main = "Residual vs. fitted")
```

#Exercise 2)

```
remove(list = ls())
par(mfrow=c(1,1))

search = read.table("search.txt",header=TRUE)
```

#1)

```
skill = c("1","2","3","4","5")
interface = c("1","2","3")
N = 15
cases = N / (length(skill)*length(interface))
for (i in 1:length(skill))
{
  for (b in 1:length(interface) )
  {
    for (c in 1:cases)
    {
      cat(skill[i],interface[b],sep = " ",fill = TRUE)
    }
  }
}
```

#2)

```
par(mfrow=c(1,2))
boxplot(search$time~search$skill) + title("Time vs. Skill boxplot")
boxplot(search$time~search$interface) + title("Time vs. Interface boxplot")
```

```
par(mfrow=c(1,1))
interaction.plot(search$skill,search$interface,search$time) + title("time vs. skill interaction")
interaction.plot(search$interface,search$skill,search$time) + title("time vs. interface")
```

#3)

```
anovasearch=lm(search$time~factor(search$skill)+factor(search$interface),data=search)
anova(anovasearch)
```

#4)

```
summary(anovasearch)
```

#5)

```
qqnorm(residuals(anovasearch),main = "QQplot of residuals")
qqline(residuals(anovasearch))
plot(fitted(anovasearch),residuals(anovasearch),main = "Residuals vs. fitted")
```

#6)

```
friedman.test(search$time,factor(search$interface),factor(search$skill))
```

#7)

```
copysearch = search
copysearch$skill = NULL
onewayanova = lm(copysearch$time~factor(copysearch$interface))
anova(onewayanova)
```

#Exercise 3)

```
remove(list = ls())
par(mfrow=c(1,1))

cream = read.table("cream.txt",header=TRUE)
```

#1)

```
anovacream=lm(cream$acidity~factor(cream$starter)+factor(cream$batch)+factor(cream$positi
on),data=cream)
anova(anovacream)
```

#2)

```
summary(anovacream)
```

#4)

```
confint(anovacream)
```

#Exercise 4)

```
remove(list = ls())
par(mfrow=c(1,1))

cow = read.table("cow.txt",header=TRUE)

#1)
cowlm = lm(cow$milk~factor(cow$per)+cow$order+factor(cow$id)+cow$treatment,data=cow)
anova(cowlm)

#2)
summary(cowlm)

#3)
cowlmer=lmer(cow$milk~cow$treatment+cow$order+factor(cow$per)+(1|factor(cow$id)),data=c
ow,REML=FALSE)
 summary(cowlmer)
 cowlmer1 =
lmer(cow$milk~cow$order+factor(cow$per)+(1|factor(cow$id)),data=cow,REML=FALSE)
 anova(cowlmer1,cowlmer)

 #4)
 t.test(cow$milk[cow$treatment=="A"],cow$milk[cow$treatment=="B"],paired=TRUE)

 #Exercise 5)

 remove(list = ls())
 par(mfrow=c(1,1))

 nauseadata = read.table("nauseatable.txt",header=TRUE)
 #1)

 medicins = c("Chlorpromazine","Pentobarbital(100mg)","Pentobarbital(150mg)")
 nausea = numeric(304)
 medicin = numeric(304)
 nausea[(sum(nauseadata$Incidence.of.no.nausea)+1):304] = 1
 y = 0
 for (i in 1:length(nauseadata$Incidence.of.no.nausea))
 {
   for (x in 1: nauseadata$Incidence.of.no.nausea[i])
   {
     medicin[x+y] = medicins[i]
   }
   y = y + nauseadata$Incidence.of.no.nausea[i]
 }
```

```
y= sum(nauseadata$Incidence.of.no.nausea)
for (i in 1:length(nauseadata$Incidence.of.Nausea))
{
  for (x in 1: nauseadata$Incidence.of.Nausea[i])
  {
    medicin[x+y] = medicins[i]
  }
  y = y + nauseadata$Incidence.of.Nausea[i]
}
naus = data.frame(nausea,medicin)

#2)
naustabs = xtabs(~naus$medicin+naus$nausea)

#3)
themed =
c("Chlorpromazine","Pentobarbital(100mg)","Pentobarbital(150mg)","Chlorpromazine","Pentoba
rbital(100mg)","Pentobarbital(150mg)")
thenaus = c("no nausea","no nausea","no nausea","nausea","nausea","nausea")
thecount = c(100,32,48,52,35,37)

newnaus = data.frame(themed,thenaus,thecount)

nauslm = lm(newnaus$thecount~newnaus$thenaus+newnaus$themed,data=newnaus)
anova(nauslm)

#4)
chisq.test(xtabs(~medicin+nausea))

# Exercise 6)

remove(list = ls())
par(mfrow=c(1,1))

airpollution = read.table("airpollution.txt",header=TRUE)

# 1)

plot(airpollution$oxidant~airpollution$day)
sat1lm=lm(airpollution$temperature~airpollution$humidity,data=airpollution)
plot(airpollution$temperature~airpollution$humidity)
summary(sat1lm)
abline(sat1lm)
```

```
pairs(~airpollution$oxidant+airpollution$day+airpollution$wind+airpollution$temperature+airpollu
tion$humidity+airpollution$insolation)

 #2)

 #step 1
 summary(lm(airpollution$oxidant~airpollution$wind,data=airpollution)) #0.5863
 summary(lm(airpollution$oxidant~airpollution$day,data=airpollution)) # 0.01093
 summary(lm(airpollution$oxidant~airpollution$temperature,data=airpollution)) #0.576
 summary(lm(airpollution$oxidant~airpollution$humidity,data=airpollution)) #0.124
 summary(lm(airpollution$oxidant~airpollution$insolation,data=airpollution)) #0.2552

 #step 2
 summary(lm(airpollution$oxidant~airpollution$wind+airpollution$temperature,data=airpollution))
#0.7773
 summary(lm(airpollution$oxidant~airpollution$wind+airpollution$insolation,data=airpollution))
#0.6613
 summary(lm(airpollution$oxidant~airpollution$wind+airpollution$humidity,data=airpollution))
#0.5913
 summary(lm(airpollution$oxidant~airpollution$wind+airpollution$day,data=airpollution))
#0.5989

 #step 3
summary(lm(airpollution$oxidant~airpollution$wind+airpollution$temperature+airpollution$insola
tion,data=airpollution)) #0.7816

summary(lm(airpollution$oxidant~airpollution$wind+airpollution$temperature+airpollution$humi
dity,data=airpollution)) #0.7964

summary(lm(airpollution$oxidant~airpollution$wind+airpollution$temperature+airpollution$day,d
ata=airpollution)) #0.7958

#3)

 #step 1

summary(lm(airpollution$oxidant~airpollution$wind+airpollution$temperature+airpollution$insola
tion+airpollution$day+airpollution$humidity,data=airpollution))

 #step 2

summary(lm(airpollution$oxidant~airpollution$wind+airpollution$temperature+airpollution$insola
tion+airpollution$humidity,data=airpollution))
```

```
 # step 3

summary(lm(airpollution$oxidant~airpollution$wind+airpollution$temperature+airpollution$humi
dity,data=airpollution))
 #step 4
 summary(lm(airpollution$oxidant~airpollution$wind+airpollution$temperature,data=airpollution))

 #5)
 finallm = lm(airpollution$oxidant~airpollution$wind+airpollution$temperature,data=airpollution)
 qqnorm(residuals(finallm),main = "QQplot Residuals")
 plot(fitted(finallm),residuals(finallm),main = "Residuals vs. Fitted")

 par(mfrow=c(1,1))

 library(plotrix)

twoord.plot(airpollution$day,airpollution$oxidant,airpollution$day,fitted(lm(airpollution$oxidant~a
irpollution$wind+airpollution$temperature,data=airpollution)))+title("oxidant vs. model")

 #Exercise 7)

 remove(list = ls())
 par(mfrow=c(1,1))

 expensescrime = read.table("expensescrime.txt",header=TRUE)


pairs(~expensescrime$expend+expensescrime$bad+expensescrime$crime+expensescrime$la
wyers+expensescrime$employ+expensescrime$pop)
 round(cor(expensescrime[,2:7]),2)

 #step 1
 summary(lm(expensescrime$expend~expensescrime$employ)) #0.954
 summary(lm(expensescrime$expend~expensescrime$bad)) #0.6964
 summary(lm(expensescrime$expend~expensescrime$crime)) #0.1119
 summary(lm(expensescrime$expend~expensescrime$lawyers)) #0.9373
 summary(lm(expensescrime$expend~expensescrime$pop)) #0.9073

 #step 2
 summary(lm(expensescrime$expend~expensescrime$employ+expensescrime$lawyers))
#0.9632
 summary(lm(expensescrime$expend~expensescrime$employ+expensescrime$pop)) #0.9543
no significant change
```

```
 summary(lm(expensescrime$expend~expensescrime$employ+expensescrime$bad)) #0.9551
no significant change
 summary(lm(expensescrime$expend~expensescrime$employ+expensescrime$crime))
#0.9551 no significant change
```

#step 3

```
summary(lm(expensescrime$expend~expensescrime$employ+expensescrime$lawyers+expen
sescrime$bad)) #0.9639 no significant change

summary(lm(expensescrime$expend~expensescrime$employ+expensescrime$lawyers+expen
sescrime$crime)) #0.9632 nochange

summary(lm(expensescrime$expend~expensescrime$employ+expensescrime$lawyers+expen
sescrime$pop)) #0.9637 no significant change

 model = lm(expensescrime$expend~expensescrime$employ+expensescrime$lawyers)
 #intercept = summary(model)$coefficients[1]
 #employ = summary(model)$coefficients[2]
 #lawyers = summary(model)$coefficients[3]

 twoord.plot(expensescrime$state,expensescrime$expend,expensescrime$state,fitted(model))
 round(cooks.distance(model),2)

 qqnorm(residuals(model),main = "QQplot Residuals")
 expensescrime = expensescrime[-c(5,8,32),]

 #extra

 anothermodel = lm(expensescrime$expend~expensescrime$employ)

twoord.plot(expensescrime$state,expensescrime$expend,expensescrime$state,fitted(anotherm
odel))
 round(cooks.distance(anothermodel),2)
 qqnorm(residuals(anothermodel))

 stepdown =
lm(expensescrime$expend~expensescrime$employ+expensescrime$bad+expensescrime$crim
e)

twoord.plot(expensescrime$state,expensescrime$expend,expensescrime$state,fitted(anotherm
odel))
```